

DOCUMENT RESUME

ED 193 241

TM 800 555

AUTHOR Guion, Robert M.; Ironson, Gail H.
TITLE Principles of Work Sample Testing. Volume I: A Non-Empirical Taxonomy of Test Uses; Volume II: Evaluation of Personnel Testing Programs; Volume III: Construction and Evaluation of Work Sample Tests; Volume IV: Generalizability.
INSTITUTION Bowling Green State Univ., Ohio.
SPONS AGENCY Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
REPORT NO ARI-TR-79-A8; ARI-TR-79-A9; ARI-TR-79-A10; ARI-TR-79-A11
PUB DATE Apr 79
CONTRACT DAHC19-77-C-0007
NOTE 263p.
EDRS PRICE MF01/PC11 Plus Postage.
DESCRIPTORS *Classification; Latent Trait Theory; *Measurement Objectives; Program Evaluation; *Psychometrics; Test Construction; Testing Programs; *Work Sample Tests
IDENTIFIERS Content Referenced Tests; Generalizability Theory

ABSTRACT

Challenges to classical psychometric theory are examined in the context of a broader range of fundamental, derived, and intuitive measurements in psychology: the challenges include content-referenced testing, latent trait theory, and generalizability theory. A taxonomy of psychological measurement is developed, based on: (1) purposes of measurement including evaluation, organizational trouble shooting, individual diagnosis, certification, prediction, and evaluation of other measurement; (2) settings in which measurements are obtained including the laboratory, institutional control, and field settings; (3) variables or attributes to be measured, including attributes of people (physiological processes, motor skills, performance variables, job knowledge, cognitive variables, aspects of personality, and attitudes), and attributes of tasks (duration or intensity of attention, hazards, degree of task structure, organizational involvement, task complexity, intrinsic feedback, skill demands, significance, and autonomy); and (4) the methods of measurement in psychology including instrumentation, direct observation and recording, records and biographical data, testing, and ratings. The challenges and the classifications are examined for special implications for work sample testing. The point is stressed that the evaluation of work sample measurement depends primarily on its acceptability as a direct description of the performance of interest. (RI)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED193241



ARI TECHNICAL REPORT
TR-79-A8

Principles of Work Sample Testing: I. A Non-Empirical Taxonomy of Test Uses

by

Robert M. Guign

BOWLING GREEN STATE UNIVERSITY
Bowling Green, Ohio 43403

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

April 1979

Contract DAHC 19-77-C-0007

Prepared for



U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

— Approved for public release; distribution unlimited.

TM 800555 1/4

**U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES**

**A Field Operating Agency under the Jurisdiction of the
Deputy Chief of Staff for Personnel**

JOSEPH ZEIDNER
Technical Director

WILLIAM L. HAUSER
Colonel, US Army
Commander

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR-79-A8	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PRINCIPLES OF WORK SAMPLE TESTING: I. A NON-EMPIRICAL TAXONOMY OF TEST USES		5. TYPE OF REPORT & PERIOD COVERED Final 15 Nov 1976 - 15 June 1978
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert M. Guion		8. CONTRACT OR GRANT NUMBER(s) DAHC19-77-C-0007
9. PERFORMING ORGANIZATION NAME AND ADDRESS Bowling Green State University Bowling Green, Ohio 43403		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q161102B74F
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, Virginia 22333		12. REPORT DATE April 1979
		13. NUMBER OF PAGES 49
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES Monitored by G. Gary Boycan, Engagement Simulation Technical Area, Army Research Institute.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Measurement theory, psychometrics, work sample testing, validity, content- referenced testing, criterion-referenced testing, latent trait theory, generalizability theory		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Challenges to classical psychometric theory are examined in the context of a broader range of fundamental, derived, and intuitive measurements in psychol- ogy; the challenges include content-referenced testing, latent trait theory, and generalizability. Psychological measurement is classified according to purpose, setting, variables, and methods of measurement. The challenges and the classi- fications are examined for special implications for work sample testing.		

(Continued)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

20. (continued)

The report, the first of four, is written for psychologists interested in psychometrics.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

ACKNOWLEDGMENTS

Conversations with many people have influenced the content of these reports, but the influence of a few of these has gone well beyond that of the usual conversation and should be acknowledged. Examples include Fumiko Samejima and Frederic Lord, who introduced me to the mysteries of latent trait theory, Robert Ebel, whose pithy comments on the obfuscation inherent in discussions of validity have inspired much of the questioning behind these reports, and Sam Messick and Mary Tenopir, who convinced me to break the content validity habit. The contributions of these people, although unintentional, are gratefully acknowledged.

Special acknowledgment is due to the help of Robert Brennan, of American College Testing, for graciously helping me come to some insights about generalizability theory, and to Gail Ironson, who has been a colleague and regular consultant on matters of mathematical and logical foundations in the three areas of challenge (content-referenced testing, latent trait theory, and generalizability theory) around which these reports have been written. Without her advice and help, and her availability for trying out ideas, they might not have been written; she shares authorship, in fact, on the paper on generalizability.

Most of all, I want to express thanks to Gary Boycan of the Army Research Institute, whose initiative has made possible the opportunity to read, sit, stare, and think about matters that have long troubled me.

TABLE OF CONTENTS

INTRODUCTION.	1
A SYNOPSIS OF MEASUREMENT THEORY.	2
KINDS OF MEASUREMENT	2
MEASUREMENT OF WORK SAMPLE PERFORMANCE	5
CONSIDERATIONS FOR A MEASUREMENT TAXONOMY.	7
PURPOSES IN PERSONNEL TESTING.	9
Evaluation of Material, Processes, or Programs.	9
Organizational Trouble Shooting	11
Individual Diagnosis.	11
Certification	12
Prediction of Future Status Events or Performance	12
Evaluation of Other Measurement	13
TYPES OF MEASUREMENT SETTINGS	14
Laboratory Settings	14
Settings of Institutional Control	15
Field Settings.	16
TYPES OF VARIABLES: ATTRIBUTES OF PEOPLE.	17
Physiological Processes	18
Motor Skills.	19
Performance Variables	19
Job Knowledge	20
Cognitive Variables	21
Aspects of Personality or Temperament	21
Attitudes	22
TYPES OF VARIABLES: ATTRIBUTES OF TASKS	23
Duration or Intensity of Attention.	23
Hazards	23
Degree of Task Structure.	23
Organizational Involvement.	24

TABLE OF CONTENTS

Task Complexity.	24
Intrinsic Feedback	25
Skill Demands.	25
Significance	26
Autonomy	26
TYPES OF MEASUREMENT METHODS.	26
Instrumentation.	27
Direct Observation and Recording	28
Records and Biographical Data.	29
Testing.	30
Ratings.	32
IMPLICATIONS OF THE CLASSIFICATIONS.	33
IMPLICATIONS OF PURPOSES.	34
IMPLICATIONS OF SETTINGS.	35
IMPLICATIONS OF PERSONAL VARIABLES.	36
IMPLICATIONS OF TASK VARIABLES.	36
IMPLICATIONS OF MEASUREMENT METHODS	37
SIMULTANEOUS IMPLICATIONS OF VARIABLES AND METHODS.	38
Reliability.	40
Logical Acceptability.	42
Acceptability of Inferences.	44
Standard-Based Interpretations	45
SUMMARY.	45
REFERENCES	48

LIST OF FIGURES

<u>Figure No.</u>		<u>Page</u>
1	Implications for type of variable and type of measurement technique for the evaluation of measurement.	39

PRINCIPLES OF WORK SAMPLE TESTING: I. A NON-EMPIRICAL TAXONOMY OF TEST USES

BRIEF

Because classical psychometric theory often seems inadequate for the development and evaluation of work sample tests, and because recent challenges to classical theory have had promising implications, the conceptual foundations of work sample testing need to be examined and clarified. This report, the first in a series of four, attempts to provide a background for that examination by considering the full scope of measurement in psychology. The purpose is to determine whether different kinds of measurement, or different circumstances of measurement, have different implications for the development and evaluation of measurement procedures.

The most fundamental approach to measurement is mathematically formal; it conforms to certain mathematically stated axioms, principally the axiom of transitivity. Fundamental measurement is expressed in formally defined units which are widely accepted throughout the scientific community. The use of such formal measurement provides rather direct descriptions, with little or no need for inferences, of the attributes of objects being measured. An example of such measurement is linear distance. One does not speak of "inferring" the length of an object through measurement, although it would be true, because the inference and the fact of the measurement are very nearly the same.

Most measurement in psychological research, and particularly the measurement described by classical psychometric theory, provides only signs from which inferences are drawn about the attributes of interest. The unit of measurement is typically the standard deviation of the distribution of a set of measurements, not a mathematically defined formal unit; traditional psychometric measurement is said, therefore, to be "norm-referenced." That is, the meaning of a score is defined relative to its position within the distribution of scores; in contrast, fundamental measurement can be applied to the single case, defining the meaning of a "score" in terms of the units of measurement in the scale used.

Three challenges to classical psychometric theory have gained in attention in recent years. One of these is a trend toward greater preference for content-referenced measurement as distinguished from norm-referenced measurement. Another is latent trait theory, which provides an analog, at least, to a mathematically formal unit of measurement. The third is generalizability theory, which seeks a more

precise understanding of the errors of measurement. All of these challenges seem to have special significance for work sample testing.

To provide a framework within which to consider classical psychometric theory and these challenges, a tentative taxonomy of psychological measurement is proposed. With it, the special issues in work sample testing can be viewed in the larger context of measurement in psychology generally. Four specific taxonomies are proposed: classifications of (a) the purposes of measurement, (b) settings in which measurements are obtained, (c) variables or attributes to be measured, and (d) the methods of measurement in psychology.

Six broad purposes of measurement are identified:

1. Evaluation of materiel, processes, or programs to permit organizational decisions to be made about them.
2. Organizational trouble shooting to identify needs for corrective actions concerning personnel units.
3. Individual diagnosis identifying strengths and weaknesses of individuals, either internally or relative to others.
4. Certification of individual proficiency or need, or levels of these, such as in the skill qualification testing program.
5. Prediction of future performance or characteristics of individuals, such as prediction for selection decisions.
6. Evaluation of other measurements, such as the use of one measurement as a criterion in the validation of another one.

Three types of measurement settings are defined. Types of variables are presented under two subheadings, attributes of people and attributes of tasks. Seven categories of the personal attributes are listed in decreasing order of objectivity of measurement, and a similar order is tentatively proposed for nine categories of task variables. Five kinds of measurement methods are identified, again in decreasing order of probable objectivity in measurement, ranging from the use of special instrumentation to the use of ratings.

Most purposes of measurement require, at least for the evaluation of measurements, at least the potential for substantial variance; argument that mastery testing, for example, should have low variance is rejected. Regardless of purpose, some form of generalizability is needed, although the diagnostic and certification purposes emphasize

the generalizability of scores while prediction requires generalizability of relationships. Regarding the categories of settings, the same statement is appropriate: generalizability across settings, either of scores or of relationships, seems universally necessary.

The implications of the joint classification of variables and of the methods for measuring them provide more diverse implications. For the most highly objective combinations, measurement must be accurate and interpretable in relation to a standard. Since work sample tests strive for objectivity, the same implications exist for them. The more subjective combinations require research into the acceptability of possible inferences as the principal form of evaluation.

INTRODUCTION

The well-established technology for aptitude testing seems inadequate for some purposes, including certification testing by work samples. In recent years, challenges to classical psychometric theory have come from many sources (e.g., Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Lord, 1952; Muir, 1977; Popham & Husek, 1969). This report, and the three that follow, will consider both classical theory and its challenges in examining (a) some special problems of work sample testing, (b) some relatively new developments in measurement, and (c) some old measurement ideas that are often ignored in psychometric discussions. By examining and clarifying the conceptual foundations of work sample testing, these papers will offer principles for the construction, use, interpretation, and evaluation of work sample tests in the broader context of general problems in the measurement of psychological variables.

The present report will identify the place of work sample testing in the context of a non-empirical taxonomy of general psychological measurement. The taxonomy will be described, and its implications for test evaluation will be presented with special emphasis on work sample testing. The second paper looks broadly at the scope of systems for evaluation of personnel testing programs. Evaluation includes psychometric concepts of validity, but it is not restricted to them.

With these broad perspectives as context, the third paper will focus explicitly on the construction and validation of work sample tests. Since the principal requirement to satisfy in work sample testing is generalizability of scores, the final paper in the series will be concerned explicitly with the problems and opportunities of different kinds of generalizability research for work samples and work sample validities.

A SYNOPSIS OF MEASUREMENT THEORY

Measurement is a characteristic scientific endeavor. No field of scientific enterprise can progress far without operationally defining, classifying, and quantifying its variables. Applied science relies especially heavily on the quantification of its subject matter. Measurement is not unique to psychological research, nor is preoccupation with an underlying theory of measurement a special prerogative of psychometrics.

Fundamental to any discussion of measurement is the fact that one does not measure objects or people; rather, one measures attributes of objects or people. Measurement implies the assignment of numbers to represent attributes according to some specified set of rules. Systems for assigning numbers can be devised for representing the weight of objects, the amount of information in a message, the amount of perceptual skill characterizing an individual, or the quality of an individual's performance. An acceptable system of measurement assigns numbers to represent only one attribute; other numbers, assigned according to other rules, can represent other attributes of the same objects, messages, people, or performance.

KINDS OF MEASUREMENT

It is useful to distinguish different kinds of measurement. Some approaches to measurement are so constructed that the numerical result in measuring an attribute of something is understood primarily with reference to the measurement system itself. In such measurement systems, the rules for assigning numbers to represent quantities are so definite, unambiguous, and widely accepted that an obtained number has an immediate and obvious descriptive meaning. Many of these intrinsically

obvious measurement processes have a foundation in clearly established natural law. Torgerson (1958) referred to such measurement as fundamental. The best examples of such fundamental measurement are physical measurements such as counting objects, weighing objects, measuring distances, and the like.

He also described derived measures, those which are derived from fundamental measurement with a similar kind of internally consistent meaning. Examples include more complex kinds of physical measurement, such as the measurement of density as a ratio of mass to volume. While these may not take their meaning in a wholly internal way, as in more nearly fundamental measurements, they take their meaning in the relationships of established scientific law relating an attribute to other attributes.

Both kinds of measurement described above are mathematically formal systems; that is, they conform to certain basic mathematical axioms such as those of transitivity or additivity, and that conformity can be demonstrated through formal mathematical proofs.

Although the most obvious examples of mathematically formal measurement are physical measurements, psychology is not without such formal systems of its own. Quite apart from the obvious behavior frequency counts (which are, of course, physical measurements of rate of occurrence), psychology has specialized fields of mathematical measurement theory such as information theory and signal detection theory. These approach measurement formally with neither interest in nor need for the conventional psychometric theory developed for traditional mental testing.

In contrast, other measurement derives meaning inferentially

more than directly descriptively. For example, formal physical measurements may be used to describe directly an attribute from which some other attribute is inferred; we speak (perhaps erroneously) of having "measured" the inferred attribute. An excellent example is the galvanic skin response; literally, one measures electrical resistance on the surface of the skin, but changes in that resistance are used for inferring changes in emotionality, and GSR is said to be a measure of emotion.

Much of psychological measurement is derived measurement, but it is statistically derived. It is not derived from statements of invariant lawfulness, as in the measurement of density; it is formally derived from statistical analyses and assumptions. The best examples of statistically derived measurement in psychology are those stemming from research in psychophysics, such as using Thurstone's Law of Comparative Judgment (Thurstone, 1959). The early history of mental testing proceeded in an analogous way; each item in a test was treated as a stimulus item, the response to which had some probability of providing an appropriate inference. The probability of an appropriate inference was increased by repeated stimulation, i.e., by using several items to make up a total measure or score. Modern computerized adaptive or tailored testing is a further example of statistically derived measurement, differing from earlier testing more in mathematical sophistication than in principle.

Statistically derived measurement is no less formal, and no less rigorous, than mathematically formal measurement derived from fundamental measurements. Most statistically derived psychological measurement has its own unique "mental unit of measurement" (Thurstone, 1959, p. 50). Whether it is the discriminial dispersion of judgments of scale separation, the variance in a set of test scores, or a

hypothetical scale for measuring latent ability, the unit of measurement in most mental measurement is the standard deviation.

Inferences may also be drawn from less formally developed measuring instruments. A fourth category of the kinds of measurement includes what can best be described as intuitive measurements. Many index numbers are established by intuitively combining a host of considerations; ad hoc tests may be constructed without prior statistical analysis but with some degree of rational thought; perhaps the best example of intuitive measurement is the ubiquitous five-point rating scale which is applied willy-nilly, without any formalisms or supporting data. Intuitive measures can be highly useful. Much of economic theory has been developed using such index numbers. As research progresses with such measurement schemes, lawful relationships are often identified which permit the development of more formal approaches to the measurement of the same variables.

MEASUREMENT OF WORK SAMPLE PERFORMANCE

Work sample testing may use all of the kinds of measurement in measuring attributes of either the work process or of the product (Shimberg, Esser, & Kruger, 1972). Intuitive scales may be used to rate or evaluate the process. Performance might be scored like paper-and-pencil tests are scored (which some forms of work sample tests actually are), using the theoretical foundations and principles for selecting items and evaluating scores used in traditional test construction and evaluation. Fundamental measures may be used to describe the product or result of performance; quality of performance can be inferred by weighing, by determining a physical breaking point, measuring conformity to tolerances, or by using other forms of fundamental, physical measurement of chosen attributes of a physical product.

Classical psychometric theory, which is but one theory among many, does not traditionally apply to, and may be inadequate for, some kinds of work sample measurement. Much mischief and confusion can result from misguided attempts to squeeze work sample testing into the same rubric used for the evaluation of inferences drawn from aptitude tests, even though many work sample variables can be appropriately handled within a conventional psychometric theory.

In considering alternatives for the evaluation of performance on work samples, it is instructive to consider challenges to traditional theory that have been offered in recent years. Perhaps the most active field of challenge is that known as content-referenced measurement, among other names, with its insistence that performance be measured not in terms of standard deviations from a sample or population mean but in terms of reaching or deviating from a specified standard level of performance (Glaser & Klaus, 1962).

Another emerging challenge to traditional psychometric theory comes from latent trait theory (Lord, 1952), or latent structure analysis (Lazarsfeld, 1950), which attempts to identify item characteristics as essentially sample-free estimates of item parameters instead of item statistics based on the sample at hand. Characteristics of a test can then be defined in terms of the characteristics of independent items comprising the test.

A third challenge comes from generalizability theory (Cronbach et al., 1972), which questions the adequacy of the traditional true score and error score division of obtained scores; it works instead to allocate the portions of total obtained score performance among various facets or conditions of measurement. In short, generalizability theory argues that it is the consistency or dependability of

measurement over varying conditions that is the important point in the evaluation of measurement.

These challenges are all relevant to the development and evaluation of work sample tests. For example, if a particular work sample is devised for a welder, and if all of the people who are administered the test perform poorly on it, there is little benefit to be derived from identifying certain people as having performed better than others; the significant statement is the content-referenced interpretation that they all performed below standard. Since work is rarely conducted under well-controlled, standard conditions, the stability of work sample performance across a reasonable range of circumstances is certainly important. The applications of latent trait theory are perhaps less obvious; it is sufficient here to note that such applications can provide a basis for standardizing interpretations of content-referenced tests over different samples of people tested in different locations or at different times.

In short, these challenges to traditional psychometric theory, and perhaps others, may lead to a newer and firmer foundation for work sample testing. It is therefore useful to examine work sample testing in context in the gamut of psychological measurement. The purpose of this examination is to determine whether different kinds of measurement, or measurement in different circumstances, have different implications for the development and evaluation of measurement procedures, particularly by work sample testing.

CONSIDERATIONS FOR A MEASUREMENT TAXONOMY

Psychological measurement does not occur as a disembodied abstraction. It occurs in the context of a broader purpose than measurement

per se, and it occurs within a broader environmental context. Purpose and setting, perhaps as much as the measurer's skill, determine what is to be measured and how one may go about it. The purposes, settings, variables, and techniques define a "gamut of psychological measurement" much more extensive than is ordinarily considered. The principal purposes of this report are (a) to suggest ways in which each of these may be classified and (b) to suggest implications of these classifications for the development and evaluation of specific approaches to measurement.

Personnel testing -- indeed, the testing movement as a whole -- occupies a relatively small portion of the total field of psychological measurement. Work sample testing, even broadly defined, occupies a correspondingly small place in the personnel testing domain. The tunnel vision of overspecialized theorizing can and does permit competent theory and practice in that branch of measurement traditionally known as psychometrics, but test theory and practice can be enriched by taking cues from a broader vision of measurement.

The implications of the different categories can sometimes focus on some kinds of descriptions of appropriate measurement, descriptions that can be expressed as simple dichotomies. The introductory remarks have emphasized one of these, the distinction between fundamental, descriptive measurement internally interpretable and more nearly intuitive, inferential measurement. Classical psychometric theory emphasizes the latter. It has also been pointed out that another possible dichotomous classification distinguishes norm-referenced from content-referenced measurement; classical theory addresses the former. Another possible dichotomy distinguishes measures of maximum performance from measures of typical performance; classical theory addresses both so long as performance can be inferred normatively.

In the sections that follow, purposes, settings, variables, and methods in measurement will be further divided, quite arbitrarily, into a number of categories. The categories are not exhaustive or particularly fine; they have not been empirically identified, nor has any attempt been made to ascertain their usefulness by determining empirically the reliability with which they can be used to classify actual measurement programs. They can, nevertheless, provide some insights into the place occupied by personnel testing, and especially by work sample testing, in the broader scheme of psychological measurement; they may also suggest principles for the evaluation of specific kinds of work sample measurement.

PURPOSES IN PERSONNEL TESTING

Nearly every use of personnel tests has in some sense a unique purpose, and each purpose has its own implications for the development and evaluation of measuring procedures. Nevertheless, some broad classes of reasonably similar purposes may be identified and examined for their special kinds of implications.

Evaluation of Materiel, Processes, or Programs. One purpose of personnel testing is to provide a dependent variable. Hypotheses that particular equipment or procedures or programs either will improve performance of personnel and should be adopted, or will have no effect or a negative effect on performance and should not be adopted, are tested in decision-oriented research. For an example, see Dobbins and Kendrick (1965) on the use of lenses in personnel detection within tropical forests.

In such circumstances, the psychological measurement of interest is usually a measure of performance. There are many ways to assess

performance; examples include ratings, counts of production or other achievements, output/input ratios of various kinds or records of production, or personnel problems over a period of time.

All of these imply some sort of work sample for program evaluation. The term is being used terribly broadly here to make the point; proficiency ratings, for example, are typically assessment of performance sampling a specified period of time, hence of a work sample of sorts. The question in assessing performance in these studies is not whether a sample of work is to be observed and evaluated but rather how effectively the sample of performance can be assessed. The first question in evaluating performance measurement is whether the sample of performance observed in the experimental setting is representative of performance in real or typical or targeted circumstances. There are also basic questions of (a) whether the performance is directly observed or only vaguely perceived (as in supervisory ratings) and (b) whether the numbers representing evaluations of performance in fact reflect irrelevant attributes of either the behavior, the worker, or the observer.

The measurement of performance in the experimental situations typical of these studies is rarely concerned with individual differences. The important unit of analysis is the group, not the individual, and the typical measure of interest is the mean performance of various experimental or control groups; "validity" is expressed as the significance of differences between these mean levels of performance. Occasionally the variance of subgroup performance will be the statistic of interest. Very rarely is the individual measure the measurement of concern in these experimental circumstances. Individual differences are usually (although improperly) treated as error variance. The reason, of course, is that the purpose of the

research is to make a decision about organizational practices or procedures, not a decision about individuals.

Organizational Trouble Shooting. A potential but not well explored use of personnel measurement is for the diagnosis or identification of organizational problems (Boyd, 1961). Measures of job satisfaction may be taken in different aspects of an organization to try to identify subgroups who may be pockets of discontent. Job knowledge tests could be given in different units to identify similar pockets of ignorance. Psychological assessment techniques may be used to identify areas of inefficiency, of inappropriate behavior, or of personnel misclassification. Most such studies are correlational in nature; such studies should attempt to maximize the relevant variances among individuals, somewhat like a magnifying glass. Other attempts to diagnose organizational problems may use quasi-experimental designs; in these studies variance within groups may be treated as error to be minimized while seeking to maximize between-group differences.

Individual Diagnosis. The term diagnosis is not restricted to clinical use. In many personnel testing uses, the purpose is to identify individual strengths and weaknesses. Sometimes the intent is to identify a person's own relative strengths and weaknesses, regardless of level. In other cases, one asks whether one individual measures up well or poorly in relation to others or, perhaps, to some standard on any given attribute. These are inferential measures; they should be chosen or constructed to yield the most acceptable and useful descriptions of the attributes assessed with minimal contamination from other attributes. A critically important issue in making comparisons is whether the measurements of different variables or from different samples can be expressed in a common metric.

Certification. A common purpose of measurement is to certify to decision-makers that individuals have levels of attributes appropriate to specific decisions. A high score on a licensing examination tells the Board of Examiners that it can decide to certify to the public that the person is competent or has certain knowledge essential to competence. The Army system of skill qualification testing is another example (Maier, Young, & Hirshfeld, 1976). Certification does not necessarily indicate anything desirable; a clinical psychologist may be required, for example, to certify to the court that a particular person is incompetent to stand trial, or to participate in his own defense, or some other form of incompetence. In personnel measurement, certification usually is intended to assure decision-makers that certain individuals have (or do not have) certain qualifications necessary for effective performance.

Certification usually implies a dichotomous decision. An individual will either be accepted for a job or for training or will not be accepted; measurement can likewise be reduced to a simple dichotomy. It should not be believed, however, that dichotomous scoring eliminates variance among people chosen; variance, like the poor, will be with us always. What is implied is that, for some uses of measurement, the amount of variance within a group may seem trivial. Measurement for certification may, therefore, be considered similar to measurement for organizational decisions or for trouble shooting; the problem may be to minimize within-group variance and maximize between-group variances.

Prediction of Future Status Events or Performance. All of the preceding categories logically imply a sort of prediction. There are, however, many purposes which may be explicitly stated in formal language as predictive hypotheses.

Where prediction is the explicit purpose, two or more measurements are involved: the measurement of the future variable -- individual status or performance or the occurrence of an event -- and the measurement of the predictor. The time element is an important part of the predictive hypothesis, and the evaluation of measurement may include an evaluation of the appropriateness of the elapsed time or other circumstances under which the measurements are taken. Descriptive measurements both at the time of prediction and the future time need evaluation. Most important is the need to evaluate not only the measurement but the tenability of the hypothesis itself.

There is nearly an infinite variety of things to predict in personnel testing. One may wish to predict whether training will be completed, level of proficiency at the conclusion of training, or proficiency or other forms of behavior at some stabilizing period after training has been completed. Each of these may call for slightly different evaluations of measurement. If one attempts to measure proficiency at the end of training, the measurement may seek to assess maximum performance capability with reference to some standard. Depending on the specific hypothesis, prediction of on-the-job proficiency may require measurement of either typical or maximum performance.

Evaluation of Other Measurement. To complete the list of purposes, it is necessary to point out that some personnel assessment is done primarily in the validation of other measurement. It may serve as a criterion measurement, as in prediction of future performance, or as the measurement of a hypothesis tested in the evaluation of construct validity.

TYPES OF MEASUREMENT SETTINGS

There is almost an infinite variety of situations in which measurements are taken. Each category below could be subdivided, some of them many times, with an increase in the precision with which settings can be described. A relatively small number of categories is used, however, because the important issue for personnel testing is the degree to which measurement is representative of "real world" situations. The categories chosen fell on a continuum ranging from artificial but highly controlled to realistic but uncontrolled situations. The higher the degree of control, the greater the loss of realism or representativeness of the research and of the measurement in it. Nevertheless, all measurement requires some degree of control or there is no standardization of measurement.

Laboratory Settings. This heading describes both actual laboratories, where full control of extraneous conditions can be maintained, and well-controlled simulations. Such control, in personnel testing, is rare except in experimental studies of human factors. Measurement in such research is usually concerned with the evaluation of a component of a system rather than the evaluation of a person or task as such. Individual proficiency in a complex skill, however, may be measured in laboratory-like simulations for certification purposes.

The emphasis is on the level of control rather than on the physical attributes of the setting. It is possible to have a highly controlled experimental study under carefully-selected field conditions. Measurement of certain attributes, such as physiological processes, may be done under conditions most nearly like those of laboratory control regardless of the physical setting in which they occur. Even within a laboratory setting, the level of control may vary; in the

study of reaction times, for example, a laboratory equipped with modern electronic apparatus can achieve a higher level of control, and therefore a greater degree of accuracy, than one where reactions are timed with a stopwatch.

The control referred to in this discussion is not experimental control over manipulations -- a major characteristic of an experiment -- but control over the measurement process itself. Without such control, attributes other than the one being measured (including attributes of different objects) are permitted to influence the measurement. With the highest levels of control, there is little influence on the obtained measurement from extraneous sources. For example, the electronic apparatus is more accurate in measuring reaction time because it does not include error due to the speed of reaction of the observer.

More accurate measurement is not necessarily better measurement. The basic problem in evaluating measurement under conditions of laboratory control is the problem of generalizability. Does measurement under the idealized, controlled conditions generalize to "real world" uncontrolled conditions? The question is an empirical one, and its importance varies with the opportunity for distortion in measurement in either artificial or clearly uncontrolled situations. What is at issue is the Brunswickian notion of representative design. Measurement taken under the relatively sterile conditions of laboratory settings may lack representativeness, and the laboratory may therefore introduce its own error by influencing the behavior or variable under study.

Settings of Institutional Control. This rather peculiar term is intended as an umbrella term covering employment offices, clinics,

training centers, and other settings in which measurement is taken under standardized (if not really controlled) conditions — conditions which include the awareness of the subject being measured that institutional decisions are going to be based on the results. Standardization implies certain conventional concerns, such as consistency in time limits, instructions, formats, etc. There are other concerns, however, that have not been handled particularly well in the psychometric literature. For example, are testing conditions standardized when the same instructions are read to all people to be tested, or when all of the people to be tested have been brought to some common level of understanding? Answers to such questions may well determine the success in minimizing unwanted influence on measurements.

Field Settings. Realistic field situations can be described on several dimensions. One might be the number of constraints on performance imposed by the environment; in some environments one may perform a wider range of tasks, or perform them with more difference in quality, than in more constraining settings. Some settings are supportive and facilitate performance of the measurement task; others are hostile environments which make it difficult to perform well. Subjectively, environments fall along a continuum ranging from pleasant to unpleasant settings, or, alternatively, motivating as opposed to inhibiting conditions.

The purposes of measurement imply the kinds of real-life conditions to which the results are expected to generalize, and they also determine whether, under those conditions, one wants to infer maximum or typical performance. It is obvious that some situations place a limiting influence on performance; conditions of measurement may need to include similar influences. Other consequences of the setting include effects on performance standards or on what may be expected

as typical performance. Field settings, in short, provide numerous sources of influence on obtained measurements. These influences across settings may not be consistent from one individual to another; scores obtained in different settings need to be compared for means, variance, and correlations to determine whether inferences from scores generalize from one setting to another.

TYPES OF VARIABLES: ATTRIBUTES OF PEOPLE

Many kinds of variables are measured in psychological research, including attributes of organizational and physical climates, architectural variables, tangible objects, social relationships and many other stimuli or behavioral outcomes. For convenience, the discussion here will be restricted to attributes of people and to attributes of the tasks they are asked to do.

The infinite variety of attributes of people have been organized below in seven categories. The categories, which certainly are not exhaustive, seem less important than the order in which they are presented. The presentation begins with a class of variables most amenable to objective measurement and concludes with variables for which little or no objectivity in measurement can be claimed.

Objectivity in psychological measurement is an elusive concept. It certainly should not be, as is commonly done, confused with a multiple-choice format. The topic will be reexamined later. For the present, modifying an earlier discussion (Guion, 1965), three considerations may facilitate objectivity in measurement:

1. Objectivity is facilitated by responses which can be empirically verified against some external standard as opposed to qualitative or evaluative responses of unverifiable substance.

2. Objectivity is facilitated by responses which are free or unconstrained, where the respondent's own preferred alternatives may be expressed, as opposed to responses which are restricted or structured by the measurement process itself (Thurstone, 1948).
3. Objectivity is facilitated by responses not easily or likely to be distorted, as opposed to responses distorted by deliberate faking, anxiety about the purposes of the testing, etc.

The common element in these is a matter of inference. Inferences can be made with more confidence, and in fact are smaller inferences, if based on responses that can be declared accurate, or are free from format constraints, or are not distorted in other ways. On the other hand, inferences are shaky indeed from faked reports of internal states or from responses which fit the format but give the respondent no option for the response that would be a better, more accurate, or more honest response.

Physiological Processes. In personnel testing, physiological variables are rarely considered except in human factors or stress research. Nevertheless, it is instructive to consider how such variables can be measured. Examples might include such diverse variables as respiratory rate or capacity; pulse, blood pressure, or other cardiovascular measures; metabolic rates or chemical concentrations; visual, auditory, or cutaneous acuity or sensitivity; and others. Measures of such variables are often fundamental or mathematically formally derived measurements. They may be measured by counting or in physical units.

It is important to be clear about the variable being measured as distinguished from the variable that might be inferred from the measurement. If we are concerned about the effect of a program of exercise on cardiovascular functioning because the purpose of the

program is to improve cardiovascular functioning, for example, we measure such functions simply as variables to be interpreted on their own terms. Frequently, however, we may be interested in the same measurements as a basis for other kinds of inference. For example, research on reactions to stressful environments may measure the same cardiovascular functions for inferences about levels of anxiety, a distinctly different type of variable.

Motor Skills. This category, too, is concerned with biological functioning; it differs in that its variables are peripheral, usually directly observable behaviors; that is, the variables do not have to be inferred from readings of instruments. Variables included in this category include dexterities, coordination, strength, and other patterns of muscular behavior.

In personnel testing, these variables are most likely to be measured as predictors in selection systems or for research on safety. This category, and the preceding one, may on occasion be measured as aspects of work sample performance; a work sample test for firefighters, for example, may consist of timing the speed with which a candidate can climb a ladder and return. An inference is involved, but it is such an easy, direct one that it is not often questioned; it is easy to infer skill in doing something but it is inadvisable to infer a lack of skill from poor performance. The assumption is that one cannot perform well without skill, but lack of skill is only one of many reasons why one would perform poorly.

Performance Variables. This is an extremely broad category, including most overt behavior. It includes, but is not limited to all measures of proficiency, speed or quality of performance, evaluations of work products, ineffective or disruptive performance, or

certain kinds of performance habits or styles -- approaches to carrying out tasks. Such variables, whether defined in terms of maximum or of typical performance, are most often used in the role of criteria or dependent variables. They may also be used as predictors or as bases for instruments certification decisions. Measures of attributes of actual behavior may be the basis for certification of proficiency or acceptability, or level of proficiency may be inferred from measurements using indirect indicators. Work samples, in most cases, are examples of performance measures, but so also are the ubiquitous ratings by supervisors. Performance is usually an objective fact, but it does not necessarily follow that its attributes can be easily or objectively measured.

Job Knowledge. Closely related to the measurement of proficiency is the measurement of the knowledge required to become proficient. Often, although sometimes erroneously, job knowledge tests are used for drawing inferences of proficiency. This use of job knowledge variables needs to be recognized as an example of a formal hypothesis; that is, it is hypothesized that a measure of test proficiency is a function of measured job knowledge. The hypothesis may often be tenable, particularly in highly complex jobs, but it usually deserves an empirical test.

Of the categories so far mentioned, this is the first in which conventional principles and methods of test construction, following classical psychometric theory, are easily used. Psychometric principles are rarely considered in the techniques for measuring physiological processes. It is true that psychometric evaluations of reliability and validity are commonly applied to measures of dexterity and coordination, and they are frequently given lip service in measuring aspects of performance. Nevertheless, this category is the first in

the list in which there are individual items that can be clearly clustered into internally consistent dimensions, the kind of items for which classical theoretical propositions, such as the Spearman-Brown formula or the theoretical foundations for definitions of parallel test, were created.

Cognitive Variables. The history of mental measurement is largely a history of the measurement of cognitive processes. It began with the measurement of intelligence (or "genius"), and much of its progress has occurred through refinements in the methods of measuring intellectual functioning. Intellectual functioning is generally considered a form of information processing, the principal preoccupation of cognitive psychology.

Typically, cognitive variables in personnel practice are measured through the use of paper-and-pencil tests. In other areas of psychology, there is evidence of discontent with this form of measurement. Lunneborg (1977) reported a series of three studies using laboratory measures of reaction time correlated with standard paper-and-pencil tests. The correlations were rather low, but the attempt to understand conventional test performance in the language of cognitive processes seemed intriguing. Cognitive variables are among the most commonly used predictors in personnel selection and classification programs; attempts to measure individual differences in these variables that utilize cognitive theory and research should be watched with interest.

Aspects of Personality or Temperament. Attempts to measure characteristics of personality have been highly varied; they include personality inventories, projective procedures ranging from ink blots to sentence completion forms, and procedures for inferring personality

characteristics from such objective data as suggestibility during an experiment, etc. More recently, attempts to assess such variables through assessment center exercises have become quite popular.

These variables differ from those under all previous headings; they are less a matter of what a person can do than of what a person will do. The emphasis is motivational, and it has no objective referent. Characteristics of personality and temperament are therefore evaluated against normative standards.

Attitudes. The measurement of attitudes involves assessing affective reactions to a wide variety of environmental characteristics. Attitude scales may be developed by scaling checklist statements, writing single item questions with graphic rating scales or other ad hoc collections of intuitively scaled response options, or by using the method of summated ratings on a series of such questions or checklist statements. The most common example of attitude measurement in personnel testing is the measurement of job satisfaction and related reactions to work and work settings.

The level of blood sugar in a given sample of blood constitutes an empirically verifiable fact; there is no way in which the level of job satisfaction of an individual in a given setting can be considered similarly verifiable. Moreover, the methods of measurement of attitudes rarely permit free responses; the responses typically are constrained by one of the formats mentioned above. Moreover, as people try to interpret the purposes of the measurement, or fear that their responses can be identified and used against them, there is a strong probability that responses will be consciously distorted. In all respects, the measurement of attitude seems to be the least objective of any of the variables in this list.

TYPES OF VARIABLES: ATTRIBUTES OF TASKS

The long history of measuring attributes of people has made it possible to organize variables describing people in a fairly coherent way. There is no comparable history in the measurement of task characteristics; although the kinds of variables to be sampled in developing work samples makes task variables extremely important to the present paper. Very briefly, nine categories of task variables can be suggested. An attempt, tentative and faltering, has been made to suggest again a rough order of objectivity or verifiability, but no definition of objectivity is offered. The earlier treatment of objectivity in terms of responses is clearly not applicable.

Duration or Intensity of Attention. Some tasks, for example that of the air traffic controller, require a constant and unwavering vigilance for prolonged periods. Other tasks require less intense attention, and even that needs to be maintained for only brief periods. Variables might differ according to the sensory modalities involved, the focus of attention, or the nature and costs of the consequences of inattention. Some of these variables may relate more to cognitive than to sensory processes, such as the number or complexity of details that must be comprehended or manipulated, or the degree to which the task demands attention to fact as opposed to attention to broad generalization.

Hazards. Physical, social, or economic risks may be intrinsic components of certain tasks. Such variables need to be considered very carefully in the development of work sample measures; a work sample designed to assess the performance of a police officer in making an arrest may, for example, be severely distorted if the sample involves simulated conditions in which the officer knows there is no chance of being shot.

Degree of Task Structure. Perhaps one of the most widely studied attributes of tasks is the degree of uncertainty (or its opposite, structure). In some tasks the outcome of performance is highly predictable. That is, one knows very clearly that doing the task in one way leads surely to specified errors, whereas performance in a different way leads to acceptable work products. In contrast, other tasks, such as artistic or craft tasks, are often carried out with very little assurance that the result will be the one intended.

Organizational Involvement. Some tasks can be done in nearly total isolation. Other tasks require a worker to receive material or ideas from other people and may also influence work of other people; examples include assembly line activities, team activities, etc. Organizational involvement may be a single variable which can be measured in terms of the number of necessary interactions with other people in an organization required to perform a task satisfactorily; alternatively, it may be analyzed into component variables as different organizational entities as the locus of involvement.

Task Complexity. Variables under this heading include the level of knowledge and skill required to carry out the task, the variety of skills demanded, the number or complexity of choices or decisions that might have to be made, the level of accountability or damages in the case of inadequate performance, or even the learning time required to perform the task effectively. It is possible to develop a work sample test using performance on relatively simple tasks as a basis for inferences about performance on a more complex task. Doing so implies, again, a hypothesized relationship between performances on the simple and complex tasks, and that hypothesis needs to be tested before its tenability is assumed.

Intrinsic Feedback. On some tasks, a worker can obtain information about how well he is doing the task as he is doing it. One who is cutting a piece of wood or metal on a lathe, for example, can periodically check the dimensions against the specifications with calipers and can evaluate his work. If one is using expendable tools, such as saw blades, and one's rate of wear or breakage is excessive relative to some standard, he can be aware of the flaw in performance without being told by an independent observer or supervisor. In other tasks, feedback about quality of performance is long delayed and may sometimes be filtered through several processes; sometimes it comes only from the subjective judgments of peers or supervisors. Work sample testing appears to be more easily directed toward tasks with opportunities for some intrinsic feedback.

One set of feedback variables may relate to the size of the task unit. The amount of time or number of cycles required to complete a unit of work, the frequency of interrupted tasks, the opportunities to set goals, the tempo or pace of the work — all of these influence the degree of feedback one gets in performing tasks. (For a discussion of these variables, see Ryan & Smith, 1954.) Once again, the importance of such variables in psychological measurement by work sample tests is that work sample tasks should have feedback properties similar to those of the work being sampled.

Skill Demands. This category includes motor, sensory, and cognitive skills (and perhaps even attitudes) that are clearly prerequisite to effective task performance. For some of these variables the task may demand quite high levels; for other variables, the level of ability demanded by the task may be much lower. These variables have special implications for work sample testing to whatever extent they change over time. Changes in the skill demands

of the job may correlate with, but should not be confused with changes in the skills of a person doing the job (Alvares & Hulin, 1973). Some changes in skills applied in the performance of a job occur with accumulated learning through experience; if this happens, the advisability of work sample testing of inexperienced people should be questioned.

Significance. This category is intended to include any variables which evaluate the importance of task outcomes. It may include the importance of the task as an influence on the performance or satisfaction of other people within the organization, it may be an element of importance for society at large, or it may involve importance for client or customers of the organization.

Autonomy. Some tasks can be performed by the worker without supervision or advice from other people; others must be done with close supervision or consultation. Autonomy is the degree to which the worker is free to do the task without the permission or advice of someone else. Another kind of autonomy might be defined as the worker's degree of discretion in making decisions; there may be different levels of discretion for different kinds of decisions about the way tasks are to be performed or the sequence to be followed in performing them. Or, autonomy might be the number of tasks that can be completed, or the period of time one may continue to work, without seeking authorization. Or, it might be the level of the worker's control over such things as pace, or sequence of activities, or quality or quality standards.

TYPES OF MEASUREMENT METHODS

The methods for measuring task attributes are related to those

for measuring the attributes of people; what differs is the nature of the inference drawn. Although a variable such as the degree of physical hazard may be determined by counting accidents, it is more often assessed by someone's judgment or perception -- a cognitive process of the observer.

All measurement of psychological attributes begins with the observation of the responses people make to specific stimulation. Differentiation among measurement techniques is necessarily based on the nature of the observational aids used and on the manner of recording responses and transforming them into measurement.

Five categories are listed. Once again, these categories are listed in the order in which they permit objectivity in measurement or, conversely, in the reverse order of the magnitude of inferential leaps necessary for the evaluation or interpretation of data. Again, as before, the categories follow this order as a matter of convenience, not as a matter of invariance.

Instrumentation. Instrumentation as used here refers to equipment, such as mechanical, electronic, or optical aids for observation. People may respond to an emotional stimulus with an increase in the moisture content of the skin surface. Except in the strongest emotional states, however, these increases may be imperceptible without the aid of galvanometers.

Many physiological responses are measured on standard polygraph instruments. Most psychological laboratories boast an array of solid state electronic circuitry for the measurement of reaction time that would have seemed like science fiction to the psychologist holding a stopwatch a mere quarter of a century ago. Sophistication in

research and sophistication in instrumentation have developed in tandem.

Instrumentation is commonplace in measuring sensory capacities or reaction times or choices. For the former, it is especially helpful in the presentation of stimulus materials, while the latter uses instrumentation to magnify, clarify, count, or record responses or characteristics of responses. The instruments may be highly sophisticated or quite simple. They may often be developed specifically for particular measurement problems. For example, Gessewein & Corrao (1971) developed special apparatus to study the possibilities of leg fractures. Their purpose was to develop a family of curves to provide designers with the means of predicting those conditions under which Naval personnel on ships would be likely to receive fractures; the variable to be measured was the force of impact as a person fell from various heights, and the technique of measurement was to have subjects drop stiff-legged onto a force gauge platform.

Instrumentation is often used in inferring work sample proficiency through measuring characteristics of the work product. In a work sample requiring the subject to make solder connections, for example, the quality of response might well be measured by measuring the conductivity of the solder connections themselves rather than by measuring responses directly. If a piece of metal is to be machined to specifications, the resulting product can be measured with anything from a ruler to laser beams to determine whether the product is within tolerances.

Direct Observation and Recording. This category is best illustrated by research in applied behavioral analysis which requires observers to count frequencies of specified behaviors. Just as

measurement techniques with instrumentation vary greatly in sophistication, so also measurement by direct observation varies greatly in the clarity, detail, and precision of instructions to observers and in the precision with which their observations may be recorded. Under many circumstances, some form of instrumentation may be a portion of the recording process. That is, the observer may make frequency counts either by making tally marks on a piece of paper or by pressing a button activating a counter.

A less exact form of measurement by observation is used in many assessment center exercises. The observers may have no specific behaviors to count; instead, they may be instructed to observe and write down "any salient behavior." At the conclusion of the exercise, the observer's record may consist both of such narrative descriptions and an evaluative rating of the behavior observed.

Records and Biographical Data. Many variables, of which attendance is perhaps the best example, are measured by frequency counts obtained not by direct observation but by examination of recorded data. Many kinds of records are maintained in most organizations. If they are maintained consistently and accurately, they provide useful data sources for the development of a variety of measures. Therein, of course, lies the rub; most systems of personnel accounting are notoriously poor. It is, however, possible to develop and maintain effective ad hoc record systems for periods of perhaps several months.

Measures of many kinds of variables may be derived from data maintained in records. For example, records may contain frequency counts of production and may also indicate periods of time away from the principal assignment when a worker cannot be expected to be

productive. By combining the two sets of data, derived measures of productivity per hour or per day can be developed. If situational factors influence daily average productivity, records can be organized so that distributions of productivity in different situations can be determined with individual production records standardized in terms of those distributions.

Records are kept in memory banks, be they file drawers, computers, or human memories. If the memory bank is in a computer, it is simply a form of storage. However, data stores in the memory of an individual is often changed in "storage" and retrieval processes. Many variables are measured by asking individuals to pull from the records of their own memories information which can be scaled, counted, or classified. It is in this context that the major difficulty in such measurement comes into clear focus: the accuracy of records must always be suspect. Records, whether from the memory of individuals or from files, suffer from variations in carefulness, in organizational procedures, in the interpretations of numbers, and in many other ways that distort their accuracy.

Testing. Personal attributes of people are most often measured by asking them questions and recording the answers to those questions; this is certainly the most common measurement technique in personnel research. Sometimes the questions are actually assignments ("Solve this problem" or "Assemble that gadget"), but the prototype of this form of measurement is the multiple-choice test item. The stimulus material is the question asked or implied in the stem, and the response is the choice of the option considered correct. If the item has a genuinely correct answer, as in an arithmetic problem, the correctness of response is highly verifiable and such tests are usually called objective. There is less verifiability of the

correctness of the response when the question deals with the subject's own typical behavior. A question might, for example, ask the subject how he prefers to spend his spare time. The optional answers might include responses such as reading a good book, going to an art museum, attending a symphony orchestra concert, or watching situational comedies on television. Many people will, of course, literally spend more time watching situation comedies if for no reason other than the ready availability of a television set; symphonies, art museums, and good books may not be as accessible. The question, of course, does not ask a factual question of how one's time is literally spent; it asks how the subject likes to spend his time, and the response to that question is not at all verifiable. Only the subject himself knows his own preferences, and he may not be sure of them. Even if he is sure, he may not be truthful. If he actually prefers situation comedies over concerts, he may nevertheless respond that he would prefer to go to a concert simply because in the testing situation he perceives this to be a more socially desirable response. Since there is no direct way to determine whether an individual has responded honestly to the question, or even whether there is a clear-cut answer, such testing is considered highly subjective.

Although the written multiple-choice question is a prototype, it is by no means the only approach to measurement by question and answer techniques. In determining how well an individual might be able to detect salient stimuli in the midst of irrelevant but pervasive stimulation, the question might be, "In which quadrant is the target stimulus?" referring to a projection on a screen. Questions in any form must be phrased appropriately. In the familiar Snellen Eye Chart, for example, the "question" may be, "Can you read the next line?" It is not appropriate for the subject to answer with a yes or no; such flippancy can be avoided by simply assigning the

reading as a task: "Now read the next line."

Ratings. When all else fails, or when energy or imagination is lacking to suggest anything better, psychological measurement consists of ratings. Some form of rating (or, more generally, subjective evaluation) is the most commonly used method of measuring performance and related behavioral variables. The basic rating system consists of a format for recording subjective evaluations of designated stimulus objects or items; the familiar graphic rating scale is only one example.

In fact, better examples involve both descriptions of observations as a basis of evaluation and the evaluation itself. The observer may note behaviors and either rate the behaviors along some designated scale or consider them in rating the ratee on a pre-determined dimension. Occasionally, the observations themselves form a rating scale. Much research in developmental psychology or in animal research requires observers to check one descriptive behavior statement observed among a list of behavior statements that have been previously scaled.

Ratings are often not based on systematic observations. Periodic efficiency reports or other methods of performance evaluation frequently consist of ratings based on the vague impressions of superiors who may never have had an opportunity to observe the subordinate's behavior directly. Research on this ubiquitous use of ratings casts considerable doubt on their utility.

Serious question may also be directed to the many forms of self-rating used in psychological measurement. Many personality inventories of a question-and-answer form require that answer to be given in terms of a scaled response. An item describing a particular form of behavior might, for example, call for response options scaled in four

steps: "very much like me," "somewhat like me," "not very much like me," "not at all like me." This, too, is a subjective judgment in which the response requires a rating along a scale. Subjects may often be given simply the assigned task to rate themselves on specific dimensions — again with the rating to be placed on a designated form.

The objectivity of ratings, or their verifiability, depends primarily on the nature of the stimulus material. Subjective ratings, or discriminations, are called for in any psychophysical measurement, such as an eye examination, yet these may be treated as relatively objective. In contrast, an instruction to rate someone on "quality of performance" is far too ambiguous to permit an interpretation of objectivity. Moreover, the objectivity of ratings depends largely on the raters' desire for objectivity; many forms of bias, ranging from the self-protection of a central tendency response bias to overt prejudice may influence recorded ratings.

IMPLICATIONS OF THE CLASSIFICATIONS

The classification schemes described in the preceding section may prove unwieldy or ambiguous if they were used to classify actual studies; it has not been empirically tried. A desirable next step would be to ask different expert judges independently to fit real examples into the categories described. If specific uses can be classified easily and reliably, support for the taxonomy would be inferred; unreliability in classification would identify needs for modification.

For the present purposes, however, no tightening of the taxonomy is necessary. These categories may not be optimal, but they are at least indicative; their implications for the construction and evaluation

of new testing programs will not differ substantially from those of an empirically modified scheme.

In this section of the report, implications will be considered first for each of the different classification schemes; they will then be considered for combinations of classifications.

IMPLICATIONS OF PURPOSES

1. For all purposes, measurement leads to decisions, and these in turn at least imply some prediction of outcomes of the decisions.
2. Work samples may be relevant for any purpose, either as dependent variables or as independent variables.
3. No class of purposes imposes restrictions to particular kinds of measurement. Although measurement of some aspect of performance is commonly intended for many of these purposes, it can be based either on fundamental descriptive measurement or on measurement requiring greater inferential leaps. Measurement in program evaluation for organizational decisions, or measurement calling for the certification of proficiencies, should in general need smaller or easier inferences than do measurements for other purposes.
4. The different purposes impose no special restrictions on the kinds of variables to be assessed; both task variables and person variables need to be assessed in meeting many of these purposes.
5. Measurement techniques which maximize variance may be used for any of the types of purposes and are highly to be desired for most.
6. Measurements taken for decisions about groups (primarily in evaluations of material, processes or groups, but sometimes in organizational trouble shooting), should provide significant group differentiation. The principle may also apply to certification (for example, to differentiate masters from nonmasters), but only if the groups are very carefully defined and if the basis for group membership is stable. These two conditions may often be impossible to satisfy.

7. Where the purpose is prediction, the evaluation of measurement must be based on how well the predictor measure correlates with a measure of the future event or state to be predicted.
8. For diagnostic or certification purposes, measurement should be evaluated by logical or statistical relationships with broader indices of proficiency or the diagnostic categories. Such evaluations can be based on the logic of content sampling, on correlations, or on experimental results.

IMPLICATIONS OF SETTINGS

1. The purposes of measurement define the set of conditions most appropriate to that measurement; this set of conditions might be termed the target conditions. In any setting differing from the target conditions, the measurement setting should be representative of the target situation in salient respects.
2. Different settings may be responsible for different contaminating variables in measurement; interpretations of the results of measurement should consider the possible distortions introduced by a particular setting.
3. Where the measurement situation differs significantly from the target situation, the generalizability of inferences from the one to the other must be assessed.
4. Measurements in laboratory settings or simulations may fail to generalize if they are over-controlled, that is, if influences expected in the target situation are not permitted to vary in the laboratory.
5. Generalizability of measurement in institutional settings is less concerned with the generalizability of scores than with the generalizability to attributes of greater institutional concern; usually, this form of generalizability is expressed as predictability.
6. When measurement is done in naturalistic or field settings, standardization requires that specific sets of conditions be used. The problem of generalizability is, in such settings, one of generalizing scores (or inferences from scores) obtained in the standard setting to other relevant settings.

IMPLICATIONS OF PERSONAL VARIABLES

1. The variables in the higher categories on this list are more likely to be tangible or directly observable and less likely to be abstract. Therefore, they can be measured more objectively, and mathematically formal methods of measurement are more likely to be available.
2. The higher the category on this list, the less appropriate is conventional norm-referenced measurement. One's pulse rate after a period of extensive exercise is not evaluated by its position in a normal distribution of pulse rates; it is evaluated with reference to a standard given the age and exercising condition of the individual whose pulse is measured.
3. Variables high in this list are likely to be evaluated primarily in terms of accuracy; accuracy is an irrelevant concern for variables low on the list. The notion of accuracy implies a well-calibrated scale of measurement, usually in units accepted by the scientific community.
4. Work sample tests are most likely to be developed to measure aspects of task performance, although in some components and under some circumstances they may measure job knowledge variables, motor skills, or physiological processes. Since work sample testing measures variables in the higher categories, these variables should be objectively measured, interpretable with reference to a priori standards, and capable of accurate measurement on a well-calibrated scale.
5. The literal measurement of one variable (e.g., skin resistance to current) may be chosen as a basis for inferences about a different variable (in the example, it might be anxiety). Such inferences imply hypotheses that need empirical verification if the inferences are to be considered valid.

IMPLICATIONS OF TASK VARIABLES

1. The variables higher on the list, in general, are associated with greater opportunity and need for objective measurement; they should be interpretable with reference to previously established standards and accuracy of measurement.

2. The identification of classes of task variables helps to define the nature of a work sample test; a first stage (and sometimes sufficient) step in the evaluation of such a test is to evaluate the degree to which it is congruent with the work being sampled on salient classes of variables.
3. Performance variables in the list of personal variables are likely to be influenced both by task variables and by settings.
4. The overall nature of a task changes with changes in settings; it follows that a major consideration in measurement of task variables is the generalizability of scores or of inferences. As a specific example, the task of cleaning a rifle in the quiet of a barracks is quite different from the task of cleaning the same rifle, with the same dirt, under fire. If a task is to be properly sampled in a work sample, the conditions of performance to be inferred should be specified. Whether performance of the task under conditions other than those specified will generalize to those conditions is an empirical question.

IMPLICATIONS OF MEASUREMENT METHODS

1. The greater the precision in specifying the response to be observed, the less the ambiguity and the greater the objectivity of measurement. Methods higher on the list promote greater specificity.
2. The more objective or fundamental the measurement technique (for example, counting frequencies), the less the inference required. Of course, one may use a fundamental measurement for an intuitive inferential jump from it; such inferences usually need empirical verification. In general, inferences based on methods high on the list are more easily verified than those based on methods low on the list.
3. Regardless of measurement technique, some form of reliability information is essential to measurement. That reliability may be the consistency assured by well-calibrated instruments, or the agreement of independent observers, or the internal consistency of scaled responses to a set of attitude items. Whatever the form of reliability of greatest concern, no measurement technique can be evaluated more

highly than the reliability permits. Reliability is rarely a sufficient evaluation, even though it is a necessary one. A set of ratings may be highly reliable because of the presence of constant errors, but the reliability is of very little value if it means no more than consistently false inferences.

4. Objectivity may be illusory. The presence of sophisticated instrumentation is not an assurance of objective measurement. The question must be asked whether the measurement obtained with such instrumentation is fundamental measurement, that is, measurement to be interpreted in terms of its own units, or whether it is a basis for a derived inference.

SIMULTANEOUS IMPLICATIONS OF VARIABLES AND METHODS

Special implications for the evaluation of measurement can come from a simultaneous consideration of the kinds of personal variables being measured and the method of measurement. In abbreviated form, condensing the classification of person attributes to five categories, the two classifications are shown in matrix form in Figure 1. The matrix is so arranged that the upper left-hand corner represents the maximum possibilities for objective measurement and the lower right-hand corner represents the maximum in necessary subjectivity.

In the extreme cases, measurement of physiological or psychomotor attributes with special measuring instruments requires only accuracy in the calibration of the measuring instruments; with accuracy, questions of reliability are moot. Concern for the generalizability of measures obtained from the situation of actual measurement to targeted situations is, of course, always a consideration in the evaluation of any measurement, but so far as the variables and methods are concerned, the closer the situation to the upper left of Figure 1, the more salient the concept of accuracy is to the evaluation of measurement. Accurate measurements are those that are most readily verifiable with

	Instru- mentation	Direct Observation	Testing	Records	Ratings
Physiological or Motor	B,D	A,B,D	A,B, (D)	B,D	A,C
Performance (Behavioral)	B,D	A,B,C,D	A,B,D	A,B,C,D	A,C
Job Knowledge	B,C,D	A,B,C, (D)	A,B, (D)	A,B,C, (D)	A,C
Cognitive or Personality Constructs	(A) ,C	A,C	A,C	A,C	A,C
Attitudes	(A) ,C	A,C	A,C	A,C	A,C

Figure 1. Implications for type of variable and type of measurement technique for the evaluation of measurement. A refers to reliability, B to acceptance of the operations, C to acceptance of inferences beyond the content, and D to interpretations relative to a standard.

reference to some standard unit of measurement such as an inch, a gram, or a count.

In the other extreme is measurement using some form of rating scale for the assessment of attitudes. There is no way in which the "accuracy" of such measurement can be verified. It is possible to obtain indices of consistency of response, but there is no way to determine whether the attitude is correctly or accurately measured. Not only are there no standard units of measurement, but there is no external referent that can be clearly said to be a better or more nearly precise statement of attitude; there is no Bureau of Standards for attitude measurement. Not even behavioral observations can be used as criteria for validating a measure of attitude; too many learned variables influence the expression or inhibition of behavior appropriate to the attitude. In a taste preference study, for example, one must simply take the subject's word for it that he evaluates one stimulus higher than the other. Thus the first kind of implication for this matrix is its influence on the permissible precision of measurement.

The above comments demonstrate an interdependence of the nature of the variable being measured and the method of measurement. Both the nature of the variable and the nature of the technique influence the saliency of different considerations in the evaluation of measurement.

Reliability. Beyond generalizability, which is universally necessary, the various cells in Figure 1 identify up to four kinds of essential evaluations for particular combinations. Cells marked with an A are those in which the first step in evaluation is an inquiry into reliability. The first step in evaluating reliability is not a computation of a reliability coefficient but an examination of the

technique of measurement itself: is the method of measurement appropriately standardized? Beyond that, the question of reliability encompasses all of the familiar concerns of equivalence, stability, and, above all, internal consistency.

In a sense, every cell in the matrix should include an A since reliability is the sine qua non of effective measurement. The cells of the upper left-hand corner, however, will have satisfied the needs for reliability automatically if the measurement can be shown to be accurate. Since accuracy has been identified as the principal consideration for this set of combinations, and since unreliable measure cannot be very accurate, then the evaluation of reliability is superfluous if accuracy is established. In all other cells, reliability often must be established as a basis for, or at least a consideration in, any other evaluative determination. Where special instruments are used, reliability may refer primarily to technical fallibility (such as trouble from poor electrical contact). Where measurement uses observers, the consistency or agreement among observers is the essential reliability. In some forms of physical or behavioral observation, the observing and recording responses may be easy enough that little or no observer error is possible or likely, and it may in such cases be unnecessary to become greatly concerned about reliability. Where observers are rating knowledge or cognition or attitudes, they are exercising their own judgments and, therefore, the likelihood of fallibility in measurement because of differences in observer judgment is very real and must be investigated. Reliability in measurement by testing is well-established in classical psychometric theory, as it is in scaling and other forms of rating. Reliability in record keeping is probably derivable from psychometric reliability; the consistency of record keeping, as well as the consistency of inferences, may be best determined by dividing records into small units of time

and comparing the data collected in different time periods. The various considerations needed for estimates of reliability will be reconsidered in the discussion of generalizability.

Reliability, it must be emphasized, is necessary in all measurement. It does not follow from that fact that reliability coefficients must always be computed. Where there is evidence of accurate measurement, it is also evidence of reliability, because there is no accuracy without reliability. Likewise, where there is evidence of validity (discussed below as "acceptability of inferences"), it is also evidence of reliability, because there is no validity without reliability. The important thing is to build the measuring instrument with care to insure maximum reliability.

A notation (A) in Figure 1 denotes particular uncertainty about effective ways to estimate reliability.

Logical Acceptability. Once reliability is established, the next evaluation concerns the acceptability of the operational definition, shown as B in Figure 1. This is largely a matter of precision in measurement; if measurement is fundamental in nature, following formal mathematical axioms, acceptance is highly probable. Statistically derived or intuitive measurements may also, however, be widely accepted, simply on the basis of the way in which the measurements are collected, if their logical foundation is persuasive enough. One issue in determining logical acceptability is whether the measurement fits its purposes in relation to the distinction between maximum and typical performance. If the purpose of measurement is to find out what people actually do in real situations, a highly controlled estimate of maximum performance cannot be accepted on logical grounds, whereas a less sophisticated form of measurement obtained under more realistic

conditions -- i.e., more representative conditions -- may be readily accepted.

The greater the objectivity in measurement, the greater the likelihood of its logical acceptability. Objectivity, it should be noted, is clearly distinguishable from construct validity, despite points of similarity. As defined in this report, objectivity depends on the degree to which the response itself is free from distortion, whereas construct validity refers to the degree to which the interpretations from the response are free from distortion by influences unrelated to a designated construct. Probably the greater the objectivity, the greater the construct validity, but the question really does not arise. What does arise is the question of whether the response is a clearly identifiable, interpretable, unambiguous response as opposed to the degree to which it is undefined and subject to varying interpretations. An inference, even from some physiological measurement, may lack construct validity even when variables are accurately measured. In medical diagnosis, for example, physicians may find symptoms easily measurable but difficult to interpret diagnostically.

Under certain circumstances, characteristics of distributions of measurements may be considered in evaluating the logical acceptability of measurement. As just one example, one may ask whether the measurement involves ceiling effects such that descriptions of individuals high on a given attribute are inaccurately obtained because of the inadequacies of the measuring technique.

Perhaps the greatest boost to the logical acceptability of a measure (well, at least its acceptability) is what is known as face validity. The term is unfairly maligned, simply because it does not in fact describe an aspect of "real" validity. Nevertheless, face

validity is of both practical and technical importance. It is practically important because it facilitates judgments of logical acceptability, at least in the middle set of cells in Figure 1. It is technically important because examinees or observers may be more appropriately motivated by measures that "look right," thus adding to the objectivity of measurement.

Acceptability of Inferences. Another set of questions refers to the acceptability of inferences extending beyond the obvious content. In the conventional way of talking about psychometric validity, most of the preceding discussion on logical acceptability referred to so-called content validity. Questions of the acceptability of inferences are, in contrast, questions of construct or of criterion-related validity. The cells marked C in Figure 1 are those where attributes can be satisfactorily inferred from the measurement only on the basis of supporting empirical evidence. In any specific case, if the nature of the measurement is inference rather than fundamental description, the psychometric concepts of the validity of the inferences are the most important aspects of evaluation. Even if the measurement ostensibly measures at a more fundamental level, inferential jumps from that level must be validated. The example given earlier should be remembered: when one uses a physiological measure not as a description of physiological functioning but as a manifestation of anxiety, the inference to be validated is the use of the measurement as an index of anxiety. The accuracy of measuring the physiological process is irrelevant. Wherever the measurement is intended to lead to an inference of attributes outside of its literal content, evidence of some form of validity, specifically criterion-related or construct validity, is essential.

The crux of classical psychometric validity is the extent to

which the variance in measurements is attributable only to the construct intended to be inferred. Insufficient validity, therefore, means that part of the variance in a set of scores is classically seen (a) as being attributable to sources of variation other than the one intended or (b) as irrelevant to the variable to be predicted.

Standard-Based Interpretations. The letter D appears in Figure 1 wherever the obtained measure should be interpretable with reference to a standard. (Some arguable cells are identified with the D in parentheses. These are generally conditions permitting substantial objectivity and in which the accuracy of measurement can be assessed. Usually, they are examples where fundamental or mathematically formal measurement is plausible.

In a sense, this could apply to all of the cells; arbitrary standards or cutting scores could be established. Cognitive test scores, for example, can be interpreted as deviations from such arbitrary points.

The intent of the designation in Figure 1, however, is somewhat different; it is intended to refer to standards defined in terms of the measurement scale, not distribution of measurements. The intent here is not so much permissive as suggestive. Wherever the purpose of measurement is certification or institutional decision-making, the aim of test specialists should be to provide measurement that can be interpreted with reference to real performance standards.

SUMMARY

Heuristic classifications of the purposes and circumstances of psychological measurement, of the variables to be measured, and of the

techniques available for such measurement have been presented. Two major conclusions should be drawn. First, conventional psychological testing is contained in only a relatively small portion of all of the classes of psychological measurement. A single-minded devotion to the principles and theory of classical psychometrics has many values, but it also has the severe disadvantage of ignoring the values of other approaches to measurement. Other approaches may be more useful where accurate descriptions rather than abstract inferences are sought; even testing for inferential purposes can be improved if the methods of obtaining the underlying descriptions are more objective and accurate.

Second, classical psychometric theory may be too narrow to use in the evaluation of measurement in some of the classes. Evaluation of measurement may include reliability and validity estimation, to be sure, but it should also include a logical evaluation of measuring techniques as operational definitions of variables, and it should seek more frequent application of the usual scientific practice of interpreting measures with reference to a priori standards.

The classifications, and the broad conclusions reached from considering them, apply to work sample testing. Work samples may be used for any of the purposes of measurement, although in these reports they are primarily considered for certification purposes. Whether the product is scored or the process of getting it, work samples fit in any kind of setting; again, however, the interest of this report is primarily in settings of institutional control. With reference to Figure 1, work samples are most likely to be tests of performance, although they may include any of the classes of variables represented by the top three rows or the classes of methods in the three columns on the left.

The common requirements for the evaluation of measurement in those nine cells are (a) assessing the logical acceptability of the measurement and (b) the possibility of interpreting scores with reference to a standard. Neither of these kinds of evaluation invokes classical concepts of validity, although evidence of validity may provide further argument in the logic supporting a measure as an operational definition of the variables measured. Moreover, conventional validity is probably necessary for job knowledge or for some performance variables if these are assessed by direct observation instead of through tests or physical instrumentation. In short, despite the fact that conventional validities may provide useful information, inferences of attributes beyond the obvious content of the work sample itself are often conspicuously absent from work sample testing and, for these cases, conventional statements of validity may be superfluous and even misleading.

This is not meant to imply that criterion-related or construct validation of inferences from work sample performance is necessarily inappropriate. The point being stressed here is that the evaluation of work sample measurement is not fundamentally an evaluation of its use in the measurement of an inferred construct or of its power to predict some external behavior; rather, a work sample is evaluated primarily on its acceptability as a direct description of the performance of interest. The demands of this kind of evaluation need careful explication.

REFERENCES

- Boyd, J. B. Interests of engineers related to turnover, selection, and management. Journal of Applied Psychology, 1961, 45, 143-149.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.
- Dobbins, D. A., & Kindrick, C. M. Jungle vision V: Evaluation of three types of lenses as aids to personnel detection in a semideciduous tropical forest. Army Tropic Test Center (Canal Zone), Report No. RR-5, December, 1965.
- Gessewin, J., & Corrao, P. Testing and modeling standing man's response to impact with applications toward predicting leg fracture to shipboard personnel. Washington, D.C.: Naval Ship Research and Development Center, Report No. NSRDC-3656, June, 1971.
- Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), Psychological principles in system development. New York: Holt, Rinehart, & Winston, 1962.
- Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.
- Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. Measurement and prediction. New York: Wiley, 1950.
- Lord, F. M. A theory of test scores. Psychometric Monograph No. 7, 1952.
- Lunneborg, C. E. Choice reaction time: What role in ability measurement? Applied Psychological Measurement, 1977, 1, 309-330.
- Maier, M. H., Young, D. L., & Hirshfeld, S. F. Implementing the skill qualification testing system. U.S. Army Research Institute, R&D Utilization Report 76-1, April, 1976.
- Melton, A. W. (Ed.), Apparatus tests. AAF Aviation Psychology Report No. 4. Washington, D.C.: Government Printing Office, 1947.
- Muir, D. E. A critique of classical test theory. Psychological Reports, 1977, 40, 383-386.
- Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Ryan, T. A., & Smith, P. C. Principles of industrial psychology. New York: Ronald, 1954.

Shimberg, B., Esser, B. F., & Druger, D. H. Occupational licensing: Practices and policies. Washington, D.C.: Public Affairs Press, 1972.

Thurstone, L. L. The Rorschach in psychological science. Journal of Abnormal and Social Psychology, 1948, 43, 471-475.

Thurstone, L. L. The measurement of values. Chicago: University of Chicago Press, 1959.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.



**ARI TECHNICAL REPORT
TR-79-A9**

Principles of Work Sample Testing: II. Evaluation of Personnel Testing Programs

by

Robert M. Guion

**BOWLING GREEN STATE UNIVERSITY
Bowling Green, Ohio 43403**

April 1979

Contract DAHC 19-77-C-0007

Prepared for



**U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333**

Approved for public release; distribution unlimited.

800355 2/4

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report, TR-79-A9	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PRINCIPLES OF WORK SAMPLE TESTING: II. EVALUATION OF PERSONNEL TESTING PROGRAMS		5. TYPE OF REPORT & PERIOD COVERED Final 15 Nov 1976 - 15 June 1978
7. AUTHOR(s) Robert M. Guion		6. PERFORMING ORG. REPORT NUMBER
8. PERFORMING ORGANIZATION NAME AND ADDRESS Bowling Green State University Bowling Green, Ohio 43403		9. CONTRACT OR GRANT NUMBER(s) DAHC19-77-C-0007
11. CONTROLLING OFFICE NAME AND ADDRESS US Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, Virginia 22333		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q161102B74F
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) --		12. REPORT DATE April 1979
		13. NUMBER OF PAGES 62
		15. SECURITY CLASS. (of this report) Unclassified
		16a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) --		
18. SUPPLEMENTARY NOTES Monitored by G. Gary Boycan, Engagement Simulation Technical Area, Army Research Institute.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Measurement theory, psychometrics, work sample testing, validity, content- referenced testing, criterion-referenced testing, latent trait theory, generalizability theory		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Suggestions are offered for increasing the objectivity of measurement in programs of personnel testing. Classical concepts of reliability and validity are reviewed. Construct validity is seen as the basic evaluation of a measuring instrument in psychology; criterion-related validity actually refers to hypoth- eses rather than to measurements, and content validity refers to test develop- ment. The major evaluation for personnel tests is less a matter of validity than of job relevance and of generalizability. Implications of latent trait theory and generalizability theory are discussed in terms of content-referenced		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

20.(continued)

testing for work samples.

This report, the second of four, is written for psychologists interested in testing and psychometrics in general.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

PRINCIPLES OF WORK SAMPLE TESTING: II. EVALUATION OF PERSONNEL TESTING PROGRAMS

BRIEF

Personnel testing should be as objective as possible. Objectivity in measurement occurs under two conditions: if the scale does not depend on who has been measured with it, and if the measures do not depend on the specific scale used. If the stimulus-response content of the test permits verifiable responses, if the format imposes no constraints on the responses, and if the responses are free from distortion, the principle of objectivity is approached.

There are aspects of the test, however, other than its stimulus-response content. Scoring procedures which are defined without respect to the content may be attached to it; inferences are often drawn going far beyond the content. The more objectively attributes can be measured, the less reaching is needed to make appropriate inferences and, therefore, the less elaborate the research needed to evaluate the measurement.

Classical concepts of reliability and validity are reviewed. Criterion-related validity is noted as concerned with inferences about other variables rather than inferences about the measure used as a predictor; criterion-related validity therefore does not evaluate the measurement per se, although it evaluates hypotheses about predictor-criterion relationships. Construct validity is seen as the essence of validity, and it is defined in terms of the proportion of total variance explainable by the construct being measured. The essence of construct validity research is disconfirmatory; that is, it is intended to consider alternative interpretations of the meaning of scores which, if supported, would disconfirm the originally proposed inferences.

Content validity is not really validity at all; it is an evaluation of the procedures of test construction, not of inferences drawn from scores. In personnel testing, the test development procedure can lead logically from definitions of a job content universe and domain to the definition of a relevant test content domain and establishment of test specifications; if the test is constructed according to those specifications, its job relevance is virtually assured. Under certain circumstances, the assurance of job relevance is all that is needed in evaluating a personnel test.

Alternatives to classical psychometric theory are examined for potential value in personnel testing, especially in work sample testing. Work sample tests are seen as being, by definition, content-referenced

tests. Latent trait theory is examined for its implications for scaling personnel tests, and the implications of generalizability research are also considered.

It is concluded that too much attention is given to classical concepts of validity and not enough to the more immediately important evaluations of job relatedness and of generalizability beyond the testing situation.

TABLE OF CONTENTS

INTRODUCTION	1
GENERAL CONSIDERATIONS IN THE EVALUATION OF TESTING PROGRAMS	1
THE TEST AS STIMULUS.	1
THE TEST AS RESPONSE.	4
THE TEST AS INFERENCE	7
THE TEST AS A TOOL FOR DECISION	8
CLASSICAL PSYCHOMETRIC THEORY: RELIABILITY	11
CLASSICAL PSYCHOMETRIC THEORY: VALIDITY	15
CRITERION-RELATED VALIDATION.	17
CONSTRUCT VALIDITY	21
CONTENT VALIDITY.	24
Job Content Universe	26
Job Content Domain	27
Test Content Universe	28
Test Content Domain.	29
The Limits of Content Sampling as Validity	29
ACCEPTANCE OF OPERATIONAL DEFINITIONS	31
Intrinsic Validity	35
Operationalism Based on Formal Structure	36
CHALLENGES TO CLASSICAL THEORY	36
CONTENT-REFERENCED MEASUREMENT	38
Work Samples as Content-Referenced Tests	40
Job Analysis	43
Assembling Test Content.	44
Scaling Test Content	44
Evaluating Content-Referenced Tests.	45
LATENT TRAIT THEORY	47
The Theoretical Foundations.	48
Uses of Latent Trait Analysis	51
Evaluation	52
GENERALIZABILITY THEORY	53

TABLE OF CONTENTS

Program Evaluation.		55
SUMMARY		56
REFERENCES		60

LIST OF FIGURES

<u>Figure No.</u>		<u>Page</u>
1	Schematic Diagram Showing Contribution to Objectivity of Three Response Dimensions (Adapted from Guion, 1965).	5
2	Venn Diagrams Relating Universes and Domains of Job and Test Content	30
3	Samples and Inferences in Work Sample Testing	42
4	Item Characteristic Curves of Three Hypothetical Items.	49

INTRODUCTION

The preceding paper in this series surveyed psychological measurement in general. It concluded with the idea that different kinds of measurement of different kinds of variables, and perhaps for different purposes, demand a different emphasis in evaluation. This paper will also be quite general, although with explicit references to the central problem of work sample testing, as it describes important considerations in the evaluation of personnel testing programs. This discussion assumes that personnel testing is best understood as taking place in settings of institutional control, even if actually done as field research, and that it covers the gamut of variables and methods of measurement.

GENERAL CONSIDERATIONS IN THE EVALUATION OF TESTING PROGRAMS

The emphasis is on the total evaluation of a total program; the evaluation of a testing program includes but should not be limited to validation. In some circumstances, conventional questions of validity may not arise at all; where valid inferences from scores must be ascertained, positive research results may be sufficient, but negative results leave many unanswered questions. A total testing program consists of offering a test under a standard circumstance as a stimulus, obtaining and scoring responses, and drawing inferences from the scores for the sake of making personnel decisions. Only the latter directly uses classical validation procedures.

THE TEST AS STIMULUS

The test content, instructions, administrative procedures, format, and the situation in which the test is administered all contribute to a stimulus complex which should be standardized; evaluation of a testing program should inquire first into the details of its standardization. Are instructions given according to clearly standard procedures? If so,

are they uniformly understood by all examinees before the test begins? If there are time limits or other constraints on performance, are they rigidly standardized and enforced, as they should be?

The principle basic to these and virtually all other questions in evaluation is straightforward: does a person's score on the test represent clearly the attribute being measured, or do other attributes of the person, the test, the procedure, or the setting in which the testing is done have some influence on the score? To the extent that irrelevant attributes influence obtained scores, a testing program is in some sense deficient.

The second line of inquiry concerns the degree to which the content, format, structure, and technique contribute to the objectivity of measurement. The term objectivity has been indiscriminantly applied in psychological measurement without much precision of meaning; when people refer to objective tests, they frequently mean multiple-choice tests. It is true that such tests may be objectively scored, but the measure becomes less than objective to the extent that the available options either constrain or suggest the responses of the examinee.

The most objective measurement is mathematically formal and is best illustrated by physical measurements. Two characteristics of such measurement make it genuinely objective: (a) the scale exists independently of the objects used in developing it, and (b) the measurement is independent of the particular instrument used for measuring. In presenting these requirements for objectivity, Wright said by way of illustration, "But when a man says he is five feet eleven inches tall, do we ask to see his yardstick?" (Wright, 1968, p. 87).

By these standards, traditional psychological testing is never objective. The measures (scores) depend on the particular sets of questions asked and on the sample of people (objects) used in item analysis, and

the meaning depends on the sample used in establishing norms. By analogy, however, some characteristics of objective measurement can be approximated in even traditional psychological testing. In objective measurement of the length of objects, for example, the measure can be verified, the instrument imposes no constraints on the results of measurement (save in the fineness of calibration), and the object itself cannot distort the measurement. By analogy, since psychological measurement is based on responses, a test is objective to the extent that its content permits responses that can be verified, and it places no constraints on the nature of the responses, and that the responses are undistorted. The reference to the multiple-choice format as "objective" suggests a further analogy in that the observer who reads the yardstick or who scores the test should not be able to distort the results.

The first contribution of the test as stimulus to objectivity is its content. A test of arithmetic skill problems can be far more objective than a measure calling for endorsements of statements of belief, partly because it is less ambiguous. In any content area, the objectivity of measurement can be enhanced if the content domain to be sampled is clearly defined and the procedures for sampling clearly specified. Clarity in defining the domain and the procedures for sampling is insurance against ambiguity. Ambiguity of content may distort responses, and it will surely lead to unreliable inferences from them.

In evaluating the test as stimulus, no special constraints need to be placed on the nature of the domain to be defined. It may be a performance domain, a domain of factual information, or a domain of approaches to measurement. For example, one may set out to construct a test of problem-solving ability. Literally dozens of problem-solving tasks may be used. One might use block designs, small assembly tasks, manipulative tasks, exercises in logical reasoning, and countless others. The domain of problem-solving is not a very unified domain. If one wanted to sample for the problem-solving test all possible kinds of problem-solving tasks,

the result would be not only an incredibly long test but one that would lack internal consistency — an important evaluative consideration. The domain, therefore, might be defined specifically in terms of the use of anagrams. Within this domain, boundaries can be established and the characteristic tasks available for sampling can be identified. One might specify that the domain will consist of seven-letter anagrams of from two to five vowels. Other specifications can be added. With the domain clearly defined, a test constructor can establish rules for sampling the domain. The rules may specify only procedures for sampling the content domain, or they may specify statistical rules for accepting items sampled. For example, if the test is to be used for conventional norm-referenced interpretations, it may be specified that item difficulties are to be within a given range and that all item-total correlations must be above some minimum value. The result of clear specifications of the test domain should be clearer meaning of scores.

These points should not be over-emphasized in an evaluation. An excellent testing program may be based on serendipitous findings using haphazardly constructed tests. Nevertheless, the final overall evaluation of the testing program is more likely to be favorable if the stimulus properties of the program have been carefully constructed.

THE TEST AS RESPONSE

The content of a test, and therefore the content of the domain sampled, is a stimulus-response content. The test is not the printed instructions or questions or assigned tasks; it is the combination of instructions and success in following them, questions and answers, or tasks and performance. Objectivity of traditional measurement from responses to stimuli, already discussed, is illustrated in Figure 1.

If the response options for the examinee are wholly open-ended, the testing content is defined in part by the content analysis of the responses

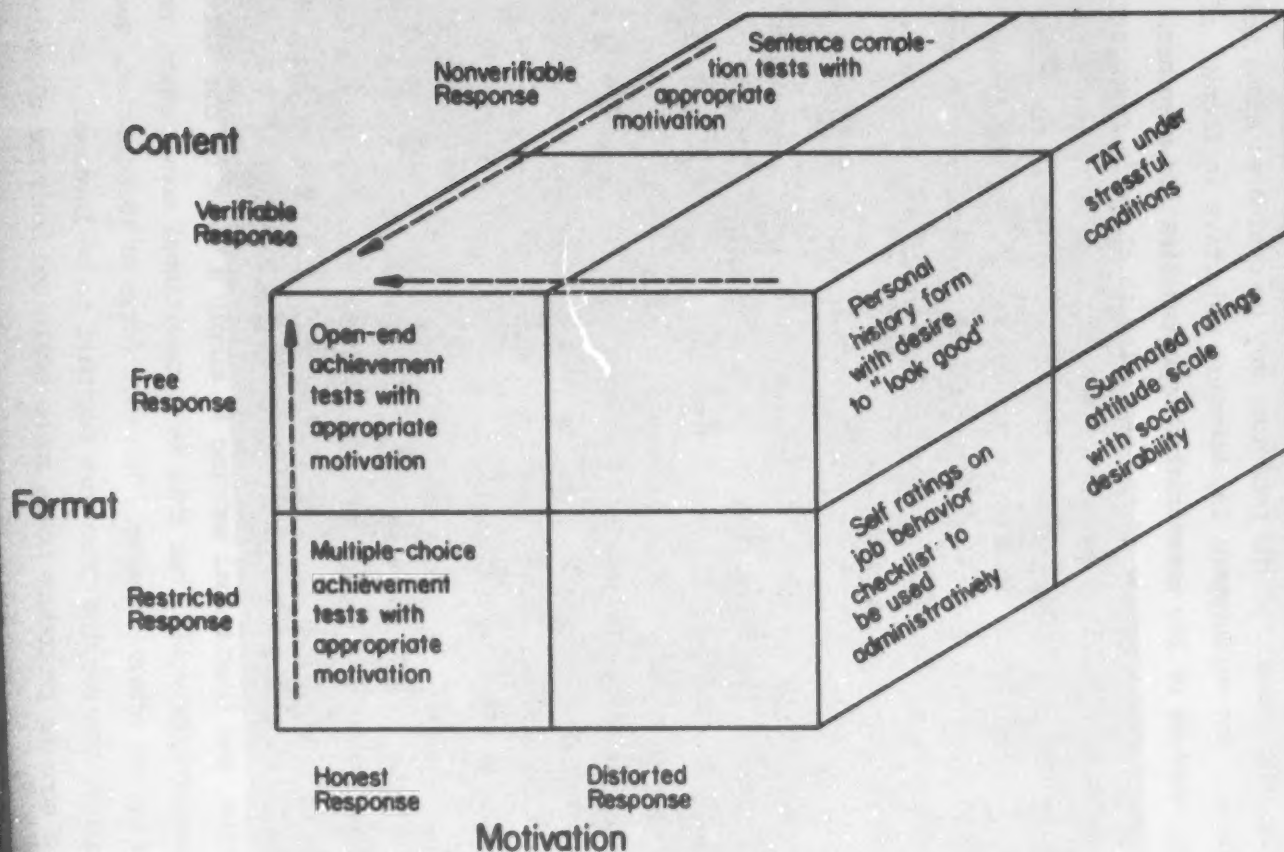


Figure 1. Schematic diagram showing contribution to objectivity of three response dimensions (adapted from Guion, 1965)

and the resulting scoring categories. Most performance tests involve open-ended responses. A work sample test, for example, consists of telling the examinee to do something. The actual responses (that is, the performance) may be observed and classified, or the consequences of the response (that is, the product of the behavior) may be evaluated along selected dimensions. The measurement is, however, objective in format, if not in scoring, because of the unrestricted opportunities for response. Objectivity may suffer (because of the necessity to classify responses), however, if the procedure permits observer or scorer characteristics to influence an obtained score intended to be a measure of an attribute of the examinee. Maximum objectivity in measurement may require an optimal tradeoff between the distortion created by artificially restricting responses and the distortion created by the unreliability or bias of observers' classification and scoring procedures.

If the form of the test is in a restricted response mode, such as multiple-choice, the definition of the test domain should include possible or plausible responses. It is a useful practice, and an indication of great care, to begin the construction of a multiple-choice test by administering the items in open-ended form to a substantial sample of people. The responses used to complete an item stem can be tallied, and a domain of potential responses can be identified with rules for selecting the correct and distracting options.

Responses to test items or tasks must yield scores if there is to be any measurement. This is one of those underwhelming, obvious kinds of statements that often seems to be overlooked. The point is that the stimulus-response content does not often include the score. In a multiple-choice test, the traditional scoring is simply a count of the number of items answered correctly, but this is a traditional convenience, not dictated by the content domain. In other forms of examination, such as work samples, the scoring procedure may have to be invented. In either case, the scoring procedure needs to be evaluated both for classical

reliability and for the possibility of contamination in the scores.

THE TEST AS INFERENCE

This heading includes the traditional concern for validity in the evaluation of testing. The topic will be examined in more detail in a subsequent section; it is sufficient here to indicate the possible variety of inferences and the evaluative questions of validity they pose.

One form of inference, according to the APA Standards (APA, AERA, and NCME, 1974), is the inference of performance in a domain based on performance on the sample. Evaluation of this sort of inference has been based on the ambiguous notion of content validity.

A different type of inference involves inferring performance on one measure from performance on a different one. Evaluation of such inferences is based on criterion-related validity.

In the third class of inferences, an individual's standing on some underlying characteristic presumably measured by the test is inferred from the score. Evaluation of such inferences is based on construct validity.

Different inferences are sought for different purposes, and therefore the emphasis on evaluating the validity of inferences is likely to differ in different testing situations. Nevertheless, it should be understood that virtually all mental measurement involves to some degree all three kinds of inference. On the anagrams test of problem-solving discussed earlier, to evaluate the inferences as valid, the evaluator must be willing to infer that performance on that set of anagrams is a good indicator of performance on any other set of anagrams from the same specified domain, he must be willing to infer that performance on the anagrams task is related to performance on something else of particular interest to the evaluator, and he must be able to infer that the performance on the

anagrams fits a network of relationships in which the scores can be interpreted in terms of problem-solving ability rather than in terms of some other characteristic such as verbal comprehension.

THE TEST AS A TOOL FOR DECISION

Most personnel testing is done to provide a basis for decisions, not primarily to measure an attribute. Evaluation of the test as a measuring instrument is important to its evaluation as a decision tool, but the two evaluations should not be confused. An excellent measure may be a poor basis for decision; a poor measure may nevertheless be the best decision tool available.

Decisions are based on predictions, either literal or implied. In personnel testing, therefore, the usual and primary evaluation of a testing program lies in the magnitude of the correlation between scores on the test and subsequent measures of the variable to be predicted. Even in situations where it is either infeasible or unnecessary to compute such a correlation coefficient, the logic of trying to maximize an implied predictive relationship remains the paramount basis for evaluating decision tools.

A principal implication of that logic is the general rule that complex performance can be predicted better with a set of predictors than with any one test. In most practical personnel prediction problems, a test battery will be devised, and some form of composite score will be computed for each person. In all discussions of test scores that follow, this composite score is as relevant as a score on a single test.

Multivariate prediction does not necessarily or uniformly imply a composite. The different variables might be arranged in some sort of sequence of decisions. Where this procedure is followed, one evaluates the testing program, and any particular test or test composite within it,

in the light of its position within the sequence. Its position in the sequence becomes another aspect of the setting in which the test is given.

The decision to be made is not an automatic consequence of the prediction. A cutting score may be set, above which individuals are selected or certified (or whatever), and it may fluctuate from time to time according to changing standards or to supply and demand. Subjective considerations may influence decisions independently of test scores. In some settings, variables that may have influenced obtained test scores may be considered by applying a mathematic correction or some sort of subjective fudge factor. Whether the decision is based solely on test scores, or whether other considerations influence the decision, the decision itself is the final step in the testing process to be evaluated. According to modern decision theory, the evaluation should be based on concepts of utility and cost effectiveness. A comparison of the costs of Type I and Type II errors should be made in evaluating the utility of the decisions.

Although the logic of prediction is almost always implied in any personnel decision, the arithmetic of prediction may be superfluous. The logical prediction frequently made, particularly when testing for a particular skill, is that a high-scoring person will perform better by using that skill if placed in a job or a training program that demands it. For example, it is almost an unarguable proposition that a person who scores high on a test of typing skill will be able to handle the typing assignments of the ordinary office. There is neither any need to compute a criterion-related correlation coefficient, nor is there much desire for it, since the extraneous factors that might inhibit performance (such as immediate conflict with the supervisor) are of little interest in the evaluation of the testing program. In these kinds of situations, the test score is interpreted on its own terms. One who types more words per minute is assumed to be able to type more words per minute than someone else. The score is its own operational definition of a skill that is prerequisite to successful performance on a job.

If the conditions of performance on the job are substantially different from the conditions of performance in the testing situation, a question of generalizability arises. To use an absurd but descriptive example, individual differences in a standardized typing test might have very little relationship to individual differences in performing the same typing task in a pitching rowboat.

The example is an extreme example of the problem of the generalizability of test scores. A permissible inference under one set of conditions may not be permissible under a quite different set of conditions. A major evaluation for many decisions is whether the generalizability of performance in the test situation to performance in the targeted conditions is a reasonable assumption. If there are to be dramatic differences in conditions, then empirical verification of generalizability yields important information.

The issue of fairness, which has been central in most discussions of personnel testing since the passage of the 1964 Civil Rights Act, should be understood as a special case of generalizability. In the situation where tests are evaluated with criterion-related correlation coefficients, the issue is one of the generalizability of the regression equation; do the constants computed for a composite of all groups apply equally well to any identifiable subgroups? In tests which are evaluated without such correlations coefficients, it may be more important to identify and evaluate the magnitude of various sources of error. Is a measure of performance on a work sample, for example, influenced by an observer's knowledge of the race or sex of the examinee? If so, to what extent? Is the task so organized that persons of unusual height have a handicap in the test situation that would not influence performance under more realistic conditions; that is, has the standardization of the test created an artificiality that influences the scores of some people unfairly because it does not exist in other conditions?

CLASSICAL PSYCHOMETRIC THEORY: RELIABILITY

It has been said that all measurement, regardless of method or attribute measured, must be reliable. It does not follow, however, that the identical ways of examining reliability apply in all cases.

The essence of an investigation of reliability is an assessment of the degree to which variance in a set of measurements may be attributed to error. Different kinds of variables, and different methods of measurement, are susceptible to different sources of error. Moreover, different methods of estimating reliability are sensitive to different sources of error (Stanley, 1971).

With many forms of measurement, one is especially interested in the stability of the description over time, that is, in errors due to instability of measurement. However, error over time is not relevant to all measurements. Blood pressure, for example, is not stable over time; it varies according to activity level, tension, etc. Yet failure to find the same blood pressure under these different conditions would never be considered to be an error of measurement; it is simply a valid reflection of the changes that occur in the attribute being measured. On the other hand, measures that are supposed to represent relatively enduring traits, such as behavioral habits or personality characteristics, should stay rather constant, at least in reasonably similar conditions, over some substantial period of time. Variations over time in measurement in these cases constitutes error. Changes in obtained measurement over time are likely to be considered sources of error for measures of personality traits, cognitive skills, motor skills, job knowledge, and most measures of performance or proficiency. It is probably inappropriate to treat change over time as a source of error in measurement for most physical or attitudinal variables.

A general principle in measurement is that one should measure one attribute at a time. The standard way of measuring, exemplified by typical tests, is to use many fallible operations to measure the same thing and accumulate observations. Thus a test will consist of many items, each with different specific content but each presumably tapping or reflecting the same fundamental attribute, -- that is, each a miniature test. The total test is a sample of observations from a homogeneous universe. Behavioral statements in rating scales constitute a similar example, as, perhaps, do pieces of information obtained through records. Obviously, observation of behavior over time can likewise be divided into "items." In short, tests can often be said to consist of component parts, each of which constitutes an independent observation of the same variable. If, however, one of these components proves to reflect an attribute other than that being measured, the inclusion of that component in the total leads to an error of measurement. An item that measures something different from the rest of the items is a contaminating item. An observation taken during a time period where a sharp noise or other distraction occurs is a contaminating observation since it reflects behavior under distraction rather than behavior under attention to the task at hand. It is conventional to refer to studies of errors in sampling the observations as estimates of internal consistency or homogeneity in measurement.

Homogeneity should be a pervasive concern in all measurement, and it is virtually assured in fundamental measurement. Tests, on the other hand, represent very small samples of nearly infinite populations of possible items and are especially susceptible to such sampling errors. To investigate these errors, it is common practice to develop and compare parallel forms. As a matter of fact, classical reliability theory assumes parallel test forms meeting rather stringent definitions. Because it is unlikely that non-test approaches to measurement will meet the requirements for parallel forms, domain sampling error in these methods may be substantially larger.

An analogous reliability problem occurs when substantially different (i.e., non-parallel) methods are used to measure the same attribute. It probably matters little whether one measures the length of a board with a flexible steel tape measure or a wooden yardstick, but in some areas of physical measurement the alternative approaches are anything but parallel. Measuring distances in cartographic analysis by triangulation is in no sense parallel to measuring with a ruler. If the two methods give different results, one or both of them may be wrong, a simple correlation to demonstrate that the methods are inconsistent is not a sufficient basis for assigning error to either one.

If observers are the instruments of measurement, there is probably a finite number (greater than one or two) of possible observers. The one or two actual observers used are samples from the universe of possible observers and, as such, may be sources of measurement error.

If two observers are used, each may contribute a unique error or measurement, and it is necessary to determine the degree to which any composite measure is subject to error in sampling observers and the degree of such error should be assessed. There is no way to estimate the error due to sampling observers if only one observer is used, just as there is no way to ascertain error attributable to a specific set of questions if only one form of the test is used. Repeated samples of observers (or tests) are required to estimate the degree to which the sampling introduces error into the measurement. Likewise, if ratings are used, some error may be due to the raters chosen, and it can only be evaluated by determining the degree of agreement between raters. Another example calls for estimating agreement among scorers of open-ended test items.

Many forms of measurement involve a subjective assignment of people or objects to scaled categories. An attempt to measure aggressive tendencies under conditions of provocation in an assessment center, for

example, might present an assessee with an anger-producing situation, and observers may be instructed to determine whether the response fits better in a category described as turning white and silent, or a category described as verbal expressions of anger, or a category described as using physical movements symbolic of attack. Such measurement poses two quite different kinds of reliability problems. One is the degree to which observers may agree on their observation of behavior; the other is the degree to which the numbers assigned to the categories fall along a reproducible scale. (The point of view taken here is that the Guttman index of reproducibility is a special case of reliability.)

Still another potential source of error can be broadly identified as a condition of measurement. The results of measurement may be different if the measurement is taken in the morning or in the late evening, it may be different if it is taken under sanitary, optimal conditions rather than in less pleasant but more realistic field conditions. Knowledge of the extent of such errors may often be useful, even if not often available.

Gross estimates of most of these sources of error can be estimated by conventional methods of estimating reliability by internal consistency coefficients, coefficients of equivalence, coefficients of stability, or coefficients of agreement (conspect reliability). However, these coefficients simply do not do a particularly clean job of separating out the components of error associated with attributes of the person, method, or setting other than the attribute measured. If one computes an internal consistency coefficient, a stability coefficient, and an equivalence coefficient, determines the proportion of variance attributable to each of the three kinds of error implied by those coefficients, and adds them up, the total estimate of error variance obtained is far greater than is realistic. A preferred approach is the generalizability analysis, or multiple facet analysis, advocated by Cronbach, Gleser, Nanda, and

Rajaratnam (1972). Using analysis of variance designs, such analysis can examine the components of error most likely to be problems in a specified testing program.

Generalizability studies seem especially useful for estimating errors in evaluating performance on a work sample. Such performance may be a function of the instructions given, the person who administers or scores the test, the activities preceding testing, and the environmental setting in which performance is being measured. One may evaluate these sources of potential error, with explicit estimates of the proportionate total variance attributable to each source, through the use of analysis of variance designs.

CLASSICAL PSYCHOMETRIC THEORY: VALIDITY

Validity refers to an evaluation of the quality of inferences drawn from test scores, qualitative summaries, judgments, or other measurements. The first point of importance in that statement is that validity is not a fact; it is an evaluation. Moreover, it is a quantitative evaluation. It is best to think of validity as expressible only in broad categories: high validity, satisfactory validity, or poor or no validity. Depending on the context, one may compare validities and say that validity in one circumstance is better, or equal to, or worse than validity in another. Since such statements do not denote precise quantities, they are not expressible with precise numbers. One should not confuse an evaluative interpretation of validity with an obtained validity coefficient. Validity is not measured; it is inferred. Although validity coefficients may be computed, the inference of validity is based on such coefficients, not equated with them.

There has been a kind of colloquial shorthand in psychometric English in which people tend to speak of the "validity of a test." Informed

people do not mean that phrase literally. It is simply a shorthand phrase referring to the evaluation of the inferences one draws from scores obtained on a test. (In this paper as in others, the shorter phrase will undoubtedly be used frequently, but there should be no misunderstanding about its meaning.) Speaking precisely, validity refers to evaluations of specific inferences that may be drawn from scores, not to evaluations of properties of tests, and there are as many validities as there are inferences to be drawn from the scores. In evaluating the total testing program, many test properties should be evaluated, such as degree of standardization, adequacy of content sampling, and the like. These properties may contribute to one's evaluation of the validity of certain inferences from scores, but they should not be confused with such inferences.

Validation refers to the processes of investigation from which the validity of certain inferences from scores may itself be inferred or evaluated. All validation procedures are in some sense empirical. Some of these procedures involve correlating test scores with other data, comparing correlations, doing experimental studies to determine differences in scores for groups differing in attributes or treatments, or evidences of procedures used in the construction of a test. Where the evidence of validity is drawn from correlations of scores with other measures, the validation does not consist simply of computing the correlation coefficient; it consists of the entire research process, including sampling of persons and of situations, the evaluations of other forms of validities of the measures used, the evaluation of the logic of the hypothesized relationship between the variables, and the purely procedural care with which data were collected. These are also empirical events, and the argument for interpreting scores on the one test as permitting valid inferences about the variable measured by the other one is supported (if at all) by the entire chain of empirical evidence, not just the correlation coefficient.

To say that certain kinds of inferences from scores are valid inferences, therefore, implies not only the empirical process of gathering data but the logical process of evaluating all of the available evidence.

Implied in the foregoing is a final observation about the nature of validity: in classical psychometric theory, validity refers to a set of scores. The evidence upon which validity may be claimed applies to the score of a single individual only if that score can be interpreted with reference to an entire set of scores. That is, in classical interpretation of scores, the individual score is considered more or less valid only if it has been previously determined that a set of scores from other individuals tested in the same way is a more or less valid set of scores. Validity is therefore defined in terms of variances; validity is the proportion of total variance relevant to the purposes of testing; irrelevant sources of variance reduce validity. A correlation coefficient describing the relationship of one measure to another is simply a means of describing the shared variance.

In short, to make judgments about the validity of the inferences one may draw from a set of scores is to make judgments about the irrelevant components in a set of scores. Earlier discussions referred to evaluations of single scores as the degree to which a score is free from reflections of attributes other than the one intended. The classical way to ascertain that freedom is to determine the level of irrelevant sources of variance. This discussion of validity in general, therefore, has reflected, without explicitly referring to them, the aspects of validity identified in the Standards for Educational and Psychological Tests (APA, et al., 1974).

CRITERION-RELATED VALIDATION

At the most directly empirical level are the criterion-related validities, predictive and concurrent. For convenience, the many reasons

for conducting criterion-related validity studies can be set in two categories: (a) to investigate the meaning that may be attached to scores on a test, that is, to identify more clearly the variable or variables measured, and (b) to investigate the utility of the scores as indicators or predictors of other variables.

The first of these grows out of the historical definition of validity as the extent to which a test measures what it "purports" the measure. If one has developed a test "purporting" to measure scholastic aptitude, then the "real" measure of aptitude is how well one does in school (Hull, 1928). School performance is then the criterion of how good the test is. That is, the correlation between scores on the test and grades in school is an index of the success of the test in measuring what it was supposed to measure. The same logic is sometimes found in modern instances in which a test of, let us say, verbal ability is correlated against supervisory ratings of verbal ability.

This kind of validation, although it involves computing a correlation between scores on the test being validated and another measure called a criterion, is better discussed under the heading of construct validity. That is, in the more conventional language of the last quarter century, such criterion-related studies are done for the purpose of verifying the interpretation of scores in terms of designated constructs.

It is an obvious outgrowth of concern for criterion-related validity that one finds that the criterion of "real" aptitude is often a variable of great importance, and the utility of the test as a predictor of that criterion becomes a matter of greater interest than the theoretical interpretation of the scores themselves. Common practice uses the term criterion-related validity primarily for those situations where one wishes to infer from a test score an individual's standing on some variable of interest that is different from the variable measured by the test. The latter variable has been called a

criterion for historical reasons, but it is usually better described as a variable analogous to the independent variable in experimental studies. The analogy is useful because, in criterion-related validities, the inference is based upon a hypothesis. That is, on a priori grounds, the test user or test developer hypothesizes that performance on the test is related to performance on some other measure, often of a different variable. Validation in such cases is less a matter of checking an intrinsic interpretation of test scores than of conducting research on the hypothesis.

In the field of personnel testing, at least for selection, the hypothesis takes the form that scores on the test can be used as indicators of potential proficiency, or some other performance variable, on a job. For example, on a given production job where each spoiled piece represents a monetary loss to the employer, scrap rate is a fundamental measure of an economic variable. With some validation, one might draw inferences about psychological variables from scrap rate (clumsiness or carelessness are competing interpretations), but this is usually not the salient point. The point is that each spoiled piece costs the organization money. If it can be shown that a particular dexterity test, or perhaps a particular test of knowledge, can predict individual scrap rates within reasonable limits of error, then the scores on the tests may be used to "infer" (more accurately, to predict) scrap rates on the job, even though the individual has not yet been trained or placed on the job. The fact that a theoretician can find an explanation for the common variance in the two sets of measurements is relatively trivial in most cases; rarely is there any attempt to interpret such criterion-related validity coefficients theoretically: what is interpreted is the value of the test as a basis for predictions of future performance. What is commonly called test validation is, therefore, best understood as an investigation of a hypothesis rather than an investigation of variables underlying scores on either predictor or criterion.

It is useful to distinguish between hypotheses that imply predictive

validity and those for which concurrent validity is appropriate. To illustrate the difference, consider the possible finding that measures of self-confidence are substantially and significantly correlated with proficiency ratings of leadership. Three independently testable hypotheses are possible: (a) that people who are self-confident become effective leaders, (b) that people who are effective leaders become self-confident, and (c) that people who are effective leaders are self-confident. The first two of these are predictive hypotheses; they predict in opposite directions. Ignoring the possibility of reciprocal causality, both of these hypotheses require predictive studies to validate them, but the design of the studies would be substantially different. In the first hypothesis, one would administer the measure of self-confidence prior to people gaining experience in leadership roles. For the second hypothesis, one would not obtain the measure of self-confidence until people have been in the leadership role long enough to establish clear and observable habits of leadership. For the third hypothesis, the two measures could be taken concurrently. The fact that very little benefit may accrue to anyone from such concurrent correlation is beside the point; the point is that the hypothesis is a different one and that in any correlational study relating them, the procedures of investigation will be different.

There has been an over-reliance on criterion-related validation in the history of personnel testing. The simplicity of the validity statement makes it very attractive, and it is often necessary for specific personnel purposes. However, things are rarely as simple as they seem, and many factors make over-reliance on a single, obtained validity coefficient questionable.

First, the conditions of a validation study are never exactly repeated. This is especially evident in the case of a predictive study, where the logic of predictive validation assumes that the conditions at the start of the study will be reasonably well matched by the conditions at the start of a new time sequence when the results of the original study are

to be applied. If a validation study extends over three or four years or more, new methods of training, new equipment, new social attitudes, new applicant characteristics, and many other new things may change the validity before the results can be put to use.

Second, the logic of criterion-related validity assumes a valid criterion. Very rarely, however, do criterion-related validity reports give any evidence of the validities of inferences drawn from the criterion measures themselves. All too often, personnel testing uses unvalidated supervisory ratings as the criterion. In many of these cases, a criterion-related validation study is probably inadvisable.

Third, the logic of criterion-related validity assumes that the sample of an applicant population used for research is truly representative and that the validity will generalize to later samples. This is almost always violated to some degree, if only through bias in attrition. Statistical procedures can, of course, provide better estimates of population validities than those provided by the biased sample, but the assumptions for these procedures often are not satisfied.

Finally, results of criterion-related validity studies, particularly those in which the predictor is a composite of several variables, are highly questionable if based on small numbers of cases. The sample size necessary to conduct a competent investigation of criterion-related validity is much larger than was earlier supposed (Schmidt, Hunter, & Urry, 1976).

CONSTRUCT VALIDITY

Despite the foregoing warnings, studies of criterion-related validity are basic in investigations of construct validity. Where the criterion is chosen because it can shed light on the intrinsic meaning of the scores being validated, such studies enable one to sharpen possible interpretations

of test scores and to choose between competing interpretations. In this context, low validities can be as helpful as high validities if they indicate what the test does not measure and thereby limit the nature of the variables legitimately inferred from the scores.

Construct validity is not a utilitarian notion. It is studied because one wishes to increase his understanding of the psychological qualities being measured by the use of a particular test. Such studies influence the degree of confidence one may have in the accuracy of descriptive inferences about the individual tested. A test is ordinarily supposed to be a measure of something; that something is an idea or concept of a variable; if sufficiently sophisticated scientifically, it is called a hypothetical construct. The latter term is intended to emphasize an idea that has been constructed as a way of organizing knowledge and experience — that is, a construct is a work of scientific imagination. As evidence accumulates about a construct, the idea may change.

The essential logic of construct validation is disconfirmatory. One does research designed to disconfirm an intended interpretation by persistently trying alternative interpretations; that is, one investigates the possibility that a variable other than the one intended to be measured (other than what a test "purports" to measure) is a better interpretation of the scores. Variance in a test intended to be used for inferring problem-solving abilities may in fact be substantially contaminated by variance due to individual differences in reading ability. Or a newly proposed construct may prove to be an old variable conventionally measured by other means. In either case, the aim of the research is to strengthen, if possible, a given interpretation of the test by showing that alternative interpretations are not feasible. Of course, if the alternative interpretation turns out to be a fairly solid one, then perhaps the originally intended interpretation is the one that is infeasible.

The notion of a hypothetical construct in its usual context is a fairly sophisticated scientific construct itself. Reference in discussions of hypothetical constructs deal with "nomological networks" of scientific lawfulness (Cronbach & Meehl, 1955). The logic and disconfirmatory emphasis of construct validation can, however, be very useful for ideas that are much less well developed scientifically. Supervisory ratings of work proficiency can, for example, be evaluated in terms of construct validity. In this case, the construct is not a highly developed creation of scientific imagination; it is a rather vague idea of proficiency on a specific job. The question is not the scientific import or sophistication of the idea, but whether proficiency is a reasonable interpretation of the variable measured by the ratings. Disconfirmatory research would consider alternative explanations. Perhaps the ratings merely measure how long ratees have been known; therefore, studies would be initiated to determine the relationship of length of acquaintance to the ratings. This is, of course, a complex question. A mere correlation between length of acquaintance and ratings may identify bias, or it may show that experience does in fact count on that job. These are competing inferences from a correlation, and the logic of construct validity requires that one attempt to evaluate them and to choose between them. In some circumstances, this might require another research study. In other circumstances, it may merely require an exercise in logic; if the job can be learned in a few days, or if proficiency is limited by external forces such as supply of material to the worker or the speed of a conveyor belt, the hypothesis that greater experience results in greater proficiency is probably silly and the correlation would disconfirm the desired interpretation of the ratings.

To say that valid inferences can be drawn about a specified construct is to say little or nothing about the utility of the measure for practical decisions. In a personnel selection situation, for example, the practical utility of the measure depends less on how well it measures a given construct

than on how well the scores will predict future performance, regardless of what or how many constructs they reflect.

As pointed out in the preceding section, in many circumstances criterion-related validity is not feasible. In these situations one exercises the logic rather than the arithmetic of predictive validity. Elsewhere (Guion, 1976), the author has used the term "the rational foundation for predictive validity" for situations in which construct validity is evaluated as part of that logic. The phrase implies that the logic of construct validation and the logic of predictive validation meet if a predictive hypothesis is very carefully developed. The steps of careful development include careful job analysis, rational inferences from the information obtained in the job analysis about the kinds of constructs that may be hypothesized as relevant to performance on the job as it would, if it could, be measured, and finally the identification of predictor variables that will validly measure those constructs. Such a logical argument pools a great deal of empirical information: the observations of the job, the group judgments involved in inferring the constructs, and the evidence of the construct validities of the predictors. None of this empirical information is necessarily expressed as validity coefficients, yet to infer that high scores on the predictors predict high performance on the job is arguably more valid under these circumstances than when a validity coefficient is obtained from an inadequate study.

CONTENT VALIDITY

Content validity is a special case of construct validity. It is likely to be emphasized in measuring knowledge or performance variables, and it is especially frequently invoked in evaluations of work samples. For that reason, it will be considered here in particular detail.

The "construct" when one speaks of content validity is more obvious

than where reference is to more abstract constructs: level of knowledge, level of skill, level of competence, or degree of mastery of a specified content or skill domain. It has been customary to speak of content validity when one wishes to infer from scores on a test reflecting the probable performance in a larger domain of which the test is a sample. We have already referred to domain sampling in sampling the kinds of items that measure a construct; the concern here is for domain sampling where the domain is more intuitively understood.

Content validity began in educational measurement as a straightforward concept which posed no special problems. An educational curriculum identifies an explicit body of knowledge and instructional objectives, and educational practice has decreed that asking a question about specific knowledge is an acceptable operation for measuring it. Therefore, if one had all possible questions about a specified curriculum content, one could obtain a universe or domain score by adding up the number of items answered correctly. When one takes a sample of all possible items from that domain, one can add up the number of items answered correctly and, from that score, infer something about the number or proportion of items that would have been answered correctly had the entire domain been used.

This account is perhaps unnecessarily glib, but the glibness gives it brevity. It is acknowledged that the best practice in sampling content domains defined by educational curricula would utilize what Cronbach (1971) called the universe of admissible operations, which identifies stimulus-response content in terms of the permissible kinds of questions and the expected kinds of responses. Nevertheless, the glibness, if that is what it is, seems defensible because the universe of admissible operations in educational testing is reasonably restricted. A combination of curriculum identification and conventional practice relieves many questions that might otherwise arise.

In personnel testing, however, the concept of content validity has been much more troublesome. The definition of a content domain has been a source of great confusion, and it is therefore necessarily difficult to define a universe of admissible operations for measuring a domain one does not clearly understand. Perhaps nowhere is the confusion better documented than in the Standards (APA, et al., 1974). In its discussion of the applicability of content validity to employment testing, that document points out that "the performance domain would need definition in terms of the objectives of measurement, restricted perhaps only to critical, most frequent, or prerequisite work behaviors." Two paragraphs further, on the same page, we read, "An employer cannot justify an employment test on grounds of content validity if he cannot demonstrate that the content universe includes all, or nearly all, important parts of the job" (p. 29).

Job Content Universe. In attempting to clarify matters, it may be useful to distinguish between the terms universe and domain and between job content and test content. We may, therefore, identify four conceptual entities: a job content universe, a job content domain, a test content universe, and a test content domain.

A comprehensive job analysis may identify all the nontrivial tasks, responsibilities, prerequisite knowledge and skill, and organizational relationships inherent in a given job, and all of this defines a job content universe.

Tasks are the things people do; job analysis need not identify trivial tasks, but it should identify the most salient activities. Responsibilities may include tasks but may also include less clearly observable activities. A teacher, for example, may be responsible for the health and safety of the children in her class. The precise activities carried out in fulfillment of that responsibility may be hard to define since they vary with changed circumstances. Prerequisite knowledge

and skill represent cognitive or motor abilities or information necessary for effective and responsible task performance. Such knowledge or skill needs to be defined unambiguously; vague trait names are not enough. "Must be able to compute means, standard deviations, correlation coefficients, and probability estimates" is a far more explicit statement than saying, "Must have knowledge of statistics."

Organizational relationships place the job in its context; they identify systems of ideas, materials, or social relationships as they influence the job; dependencies that may exist in sequences or task performance, and the degree to which people in other jobs must depend on the incumbent in the job being analyzed in doing their job. Both the organizational relationships and the responsibilities describe not only the content of the job but the content of the consequences of the performance of a job.

If the job is at all complex, it would be either impossible or absurdly impractical to try to develop a work sample test to match that total job content universe, it might be necessary to carry out the full training, to provide experience, and to observe performance on the actual job for a period of time. If one's purpose were selection, this would be absurdly impractical.

Job Content Domain. In practice, one identifies a portion of the job content universe for the purposes of testing. In a stenographic job, for example, the portion of the universe most salient in selection or performance evaluation might be restricted to those aspects involving typing. From the job analysis, one could identify the tasks, responsibilities, and the prerequisite skill (such as spelling) associated with typing; with these restricted elements, and ignoring other aspects of the total job content universe, one can define a task content domain. In this sense, the word domain is being used as a sample (and not necessarily a representative one) of the content implied by the word universe.

Test Content Universe. Performing a job and taking a test are not identical activities, even if the component elements are identical. To continue with the stenographic example, typing mailable letters from dictation on a real job involves interruptions, knowledge of the idiosyncracies of the person who has dictated the letters, interruptions by telephone calls or requests for materials from files, etc. Typing from the same dictated material in a test situation involves typing under the anxiety created by the testing and its peculiar motivational characteristics, in standard conditions such that any distractions are built into the exercise and are standardized for all people, and using material dictated by an unfamiliar voice. To the best of this writer's knowledge, no one has ever developed a typing test that is a genuine work sample in the sense of duplicating actual circumstances, distractions, and snide comments on the dictation tape -- nor has he encountered anyone who would advocate it.

Instead, one defines from the job content domain a universe of possible operations for the development of a test. The test content universe, therefore, consists of all of the tasks that might be assigned, all of the conditions that might be imposed, and all of the procedures for observing and recording responses that might be used in the development of the content sample. The test content universe is, again, a sample of the job content domain. But it is more than that; it includes elements that are not part of the job content domain since the latter probably includes no information about procedures for observing and recording behavior on assigned tasks. This would be particularly true if the operations decided upon consisted of a series of questions about the reasons for certain procedures in carrying out a task. One would virtually never include such question-and-answer exercises as a part of the actual job, but they can be quite useful in testing people to determine their qualifications for the job.

This is a crucial point in the total chain of argument. In many kinds

of work sample testing, psychometric considerations require the inclusion of non-job components in defining a test content domain; otherwise, there may be no measurement of anything. Such added operations may involve ratings by observers, counting (and perhaps weighting) responses to questions, or carrying out physical measurements and inspection of products that go beyond those encountered in the actual job itself but are necessary foundations for measurement.

Test Content Domain. The test content domain is a sample of a test content universe, and it defines the actual specifications for test construction. Again, the test content domain is not necessarily a representative sample of the test content universe. Questions of practicality and of relative importance must assuredly enter into the judgments defining a test content domain.

There remains, then, the actual construction of the test.

The Limits of Content Sampling as Validity. The foregoing sequence, which is illustrated by Figure 2, is not necessary as a detailed procedure, but the four-step process of domain definition is useful for clarifying the relationships of job and test domains and for reconciling the contradictory statements in the Standards.

It should be clear that what has been called content validity is quite different from all other forms of validity. As a matter of fact, the term should not be used since it can only cause confusion. The term validity refers, as has been pointed out, to an evaluation of the inferences that can be made from scores. If the inference to be drawn from a score on a content sample is to be an inference about performance on an actual job, it is drawn at the end of inferential leaps, in any one of which there can be a serious misstep. The crucial chance for misstep is in the definition of a test content universe; it is here that a system of scoring (or its basis) is invented, and that system of scoring is

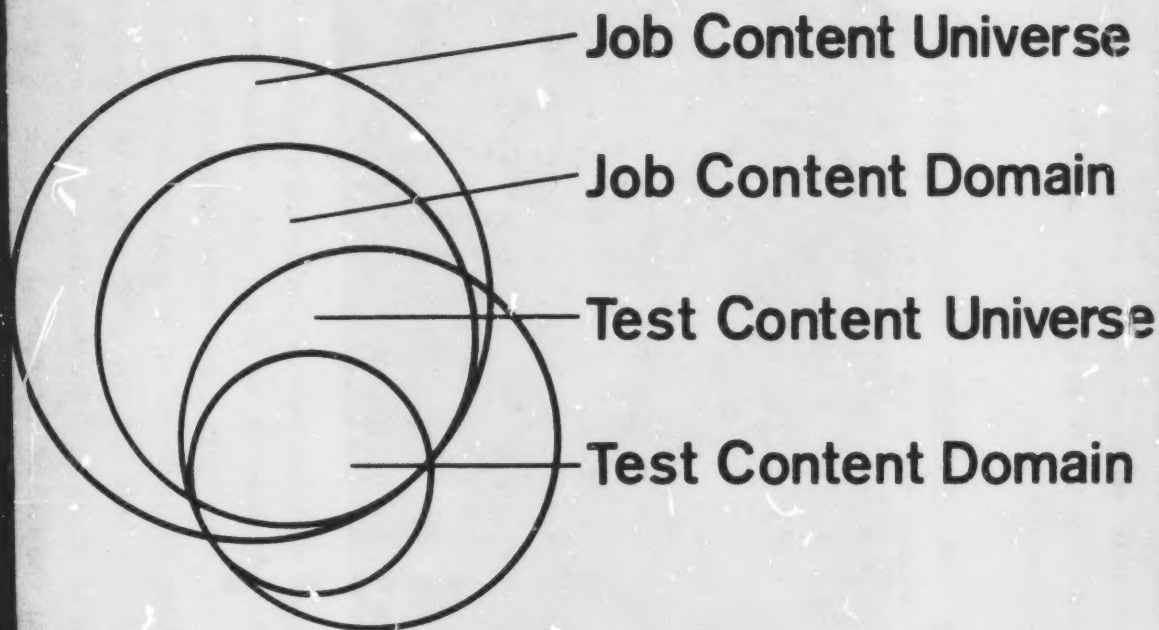


Figure 2. Venn diagrams relating universes and domains of job and test content

rarely if ever a component of the actual job content domain. Moreover, the scoring system is subject to contamination, just as is the scoring of any other test. That is, the obtained score an individual makes may reflect the attribute one wishes to infer, ability to do the job, but it may also reflect a variety of contaminations such as anxiety, ability to comprehend the verbal instructions, or perceptual skills in seeing cues for scoring enabling perceptive or test-wise people to make better scores than others.

All of this has a familiar ring after the earlier discussion of construct validity. All of the other possible components of a score represent the contaminations which construct validation, in its commitment to disconfirmatory research, is designed to investigate. To repeat: content validity is a special case of construct validity (Messick, 1975; Tenopir, 1977).

ACCEPTANCE OF OPERATIONAL DEFINITIONS

"Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few. Test validation, in fact, is widely regarded as the least satisfactory aspect of test development.... It is the purpose of this paper to develop an alternative explanation of the problem, and to propose an alternative solution. The basic difficulty in validating many tests arises, we believe, not from inadequate criteria but from logical and operational limitations of the concept of validity itself. We are persuaded that faster progress will be made toward better educational and psychological tests if validity is given a much more specific and restricted definition than is usually the case, and if it is no longer regarded as the supremely important quality of a remittal test" (Ebel, 1961, p. 640).

With these words, Ebel began a critique of the concept of validity as a major basis for evaluating tests. Many of the comments made in that paper are still highly applicable; people still tend to think of validity in terms of "real" traits, they still accept criterion measures

that have little if anything to do with the attributes being measured (and do not recognize that in doing so they have formed an external hypothesis), and the concept of validity is still far too broad to have scientific utility. Alternatives include the evaluation of reliability, normative data, the importance of the knowledge or abilities required by a test, convenience in the use of the test, and, most of all, meaningfulness.

Meaningfulness was also the primary yardstick for evaluation proposed by Messick (1975); his concept of meaningfulness, however, turns out to be equivalent to the concept of construct validity. Ebel, but not Messick, would evaluate a test simply as an operational definition of an attribute to be measured; the operations provide the meaning.

This writer takes the position that operational definitions of the attributes to be measured can, under certain circumstances, provide both a necessary and a sufficient evaluation of the scores obtained by using it; that is, under certain circumstances, no statement of validity is needed. It is operationalism, not validation, that provides the meaning for fundamental measurement of physical properties of length, time, or weight. As pointed out in the taxonomy of measurement, for measurement of such variables as these, one asks not whether the measurements are valid but whether they are accurate.

Some psychological measurement can also be defended as meaningful because of the operations involved in the measurement without recourse to the psychology's unique demand for validating the inferences from the scores. Operationalism does not always eliminate concern for validating inferences; in fact, it is sufficient only in relatively restricted cases (Ebel, 1956, 1961; Tenopir, 1977). In Tenopir's terms, there are some constructs for which the content of the measurement, i.e., the operational definition, is a sufficient evaluation. With reference to the taxonomy of variables describing attributes of people, it would appear that these

constructs would include certain physical attributes, psychomotor skills, task proficiencies, and, with a caveat, measures of job knowledge. (The caveat is that scores on job knowledge tests may be unduly influenced by reading abilities having little to do with the actual level of knowledge.)

For these constructs, at least in part, it would seem possible to evaluate the job relevance and meaningfulness of a personnel testing program on the basis of the operations alone. In a combination of two other publications (Guion, 1977, in press), the writer has presented a list of six requirements which, if met, constitute a sufficient evaluation of the use of a test so that issues of validity need not arise. With some modifications to fit the present context, and with emphasis on personnel testing and judgment of job relatedness, these will be reproduced here.

First, the content domain must consist of behavior the meaning of which is generally accepted. At the risk of sounding like Gertrude Stein, we can say that doing something (like driving a car) is generally accepted as evidence of the ability to do it. If a person reads a passage, it means that he can read the passage; if he does not read the passage, it may not mean an inability to read it (Messick, 1975), but it certainly means that he did not. In such examples, the meaning of the behavior is obvious; it requires no great inferential leap to interpret or to draw inferences from the behavior samples.

Second, both the test content domain and the job content domain should be unambiguously defined. The domains should be defined well enough that people who disagree on the definition can nevertheless agree on whether a particular task or statement or item belongs in or out of the domain. In the present age of litigation, agreements on the definition of a content domain are always tenuous. The amount of agreement needed does not depend on nailing down, in very precise language, every conceivable component of a domain. It is enough that the boundaries of the domain are sufficiently well established for agreement among reasonable and knowledgeable people.

Third, the test content domain must be relevant to the job content domain. The question of relevance is again a matter of judgment, and judgment requires some evidence of agreement. In originally presenting this third condition, the lack of a measure of the degree of agreement of the domains was deplored; it now seems that the extent of agreement among qualified judges that the two are comparable is sufficient.

Fourth, qualified judges must also agree that the test content domain has been adequately sampled. The need to define what is meant by qualified judges is particularly strong in this condition. From the point of view of personnel testing, the best qualified judges are usually people who have done the job in question or who have supervised the performance of that job. The required level of agreement would appear to be minimally that necessary to avoid conflict. Disagreements differ qualitatively. Some qualified judges will disagree on semantic grounds; others may disagree because of fundamental differences in value systems. The disagreement between plaintiff and defendant is a serious level of disagreement; the disagreement between one who would suggest a slight change in wording and one who prefers the existing wording is not a profound disagreement and need not be taken seriously in evaluating domain sampling. The question, therefore, is whether there is a consensus (a majority view) and whether there is a reasonable freedom from dissatisfaction with the consensus on the part of most qualified judges. This requirement holds for defining the boundaries of a content domain, for judging the relevance of a test content domain to a job content domain, and for judging the adequacy of the sampling of the test content domain.

Fifth, the response portion of the testing must be reliably observed and evaluated. In the original presentation of this point, it was said, "This does not refer to internal consistency, of course" (Guion, 1977, p. 7). The phrase "of course" is now regretted. At the very least, any measurement should have some degree of functional unity; if there is not even enough internal consistency for significant correlations to exist

between the component parts of a content sample, then the score of the content sample should be subdivided into reasonably internally consistent components. This comment, it should be pointed out, is a necessary consequence of saying that what has passed for content validity is in fact a special case of construct validity; the first requirement of construct validity is internal consistency.

A more important implication of this requirement is that observers who record observations must agree reasonably well on what they have seen. If the behavior to be observed is not defined well enough to permit inter-observer agreement, it violates the first condition of an operational definition based on content sampling.

Sixth, the method of scoring the content sample must be generally accepted by qualified judges as relatively free from contaminants reflecting irrelevant attributes of examinees or attributes of observers or materials. This implies no stringent demands for agreement among the judges. If there is a serious suggestion of contamination from judges who have made the previous judgments, some study inquiring into the construct validity of the scores may be necessary.

Intrinsic Validity. A different approach to operationalism can be drawn from a parallel to the concept of intrinsic validity (Gulliksen, 1950), another way in which the meaningfulness of an operational definition can be known by its outcomes. For example, if an examinee is coached to take the test, and coaching for the test improves both test performance and performance on the job, then scores on the test are intrinsically related to performance on the job. The investigation of this relationship is, of course, an empirical investigation; it does not rest upon the consensus of qualified judges. Nevertheless, it is only remotely related to its closest cousin among the validities, criterion-related validity. For the test to be accepted as an operational definition, under this heading, not only must a correlation between

test performance and job performance be obtained, but it must not be lost as a consequence of coaching.

Operationalism Based on Formal Structure. If work sample performance is to be evaluated by evaluating the product of that performance, and the product is a tangible object, then the measurement may consist of measuring weight, conductivity of solder connection, the amount of stress needed to break a weld, or similar physical measurement. Such measurements are formal, fundamental measurements and they need no justification by recourse to notions of validity.

The logic of formal measurement could be extended to some other areas of psychological measurement. Two possibilities seem worth mentioning which, if tests could be successfully constructed by these methods, would provide formal measurement that should be accepted without any concern for notions of validity. One of these uses Guttman scaling for content-referenced interpretations of scores; the other applies latent trait theory. These will be discussed in detail in the next section.

CHALLENGES TO CLASSICAL THEORY

Classical psychometric theory has its origin in the study of individual differences. This study requires maximum distinctions between individuals, that is, maximum variances within groups. All of classical theory is based upon variance and upon the subdivision of variance into systematic and error sources. A test is said to be reliable, for example, to the extent that the variance in a set of scores obtained through its use is free from random error of variance. In its broadest sense, validity is likewise defined as the extent to which the variance in a set of scores is relevant to the purposes of measurement. In test construction, the best items are those in which there is a good match between item variance and total test score variance. The unit of measurement in mental testing is the standard deviation, and the basis for interpreting

test scores is the relationship of one individual to another in distribution.

In short, the emphasis has been on relative measurement rather than on anything fundamental or absolute. The contributions of classical psychometric theory have been substantial, but they have led to some peculiar phenomena. For example, grades in a course of study such as physical education may be based not on the number of pushups one can do or the distance one can swim, but rather upon how many pushups or how many laps one can do in comparison to others in the class. Special characteristics of the class do not enter into the standard evaluation of performance using classical theory.

The illustration points out three objections that have been leveled against the use of classical psychometric theory for many forms of measurement in psychology in general and in personnel testing in particular:

1. The evaluation of measurement and of the interpretation of individual scores depends on the unique characteristics of the sample of people and the sample of items studied in the construction and standardization of the test.
2. Classical interpretations of scores provide no standard for the interpretation of an individual score beyond its relative position within the distribution of scores in the sample of people studied. If the distribution as a whole is quite high, a low score within that distribution is treated as a poor score, even if in some absolute term it would be considered high. Even the techniques for estimating true scores are based upon sample distribution; estimation of a so-called true score is simply a device for acknowledging the fallibility or unreliability of measurement. It does not take into account the relationship of that estimated true score to any standard of measurement.
3. Classical measurement theory offers no definition of the limits of the usefulness of the test or of the degree to which the classical statements of validity, reliability, or norms may be generalized. No sample is ever precisely like the sample upon which norm tables have been built, but those tables are

consistently used for interpreting the scores of people not in that sample. To what extent do these interpretations apply to people who are different from the original sample in certain ways? To what extent can the standardized interpretations of scores as norms be applied to different sets of conditions? Such questions have no answer in classical psychometric theory.

Three challenges to classical psychometric theory can be identified and discussed as potential solutions to this set of problems: content-referenced measurement, latent trait theory, and generalizability theory. In addition, another "challenge" is based on the fact that psychometric theory evaluates only inferences from scores, not the effects of the uses of such inferences. Program evaluation will be briefly mentioned in this context.

CONTENT-REFERENCED MEASUREMENT

The term content-referenced measurement will be used here to apply to any measurement technique developed explicitly to interpret scores relative to some sort of standard. The nature of the standard may vary; it might be a relatively precise point, perhaps with very tight tolerances, as in measuring machined work products. It might be a much more diffuse range of measurements, as in defining a range of satisfactory "scores" in physiological measurements associated with health. It might be an arbitrary cutting point, above which some people are selected and others rejected. However it is defined (and the defining of a standard identifies one of the problems with the relevant literature), that definition results in interpretations of scores relative to the internal structure or content of the measuring instrument rather than to a distribution of obtained measures. Whatever it is, and there is much debate over its precise nature, the one point to be emphasized is that content-referenced measurement is not norm-referenced measurement!

In keeping with the APA Standards, the term chosen here is content-

referenced measurement in preference to the more common term, criterion-referenced measurement. In most problems in educational measurement, the distinction between the two terms may be trivial enough to explain why, despite the preference in the Standards, the former has not been adopted; moreover, the term, criterion-referenced, has been so widely accepted in educational circles that there is a very real problem in attempting to change it (Hambleton, Swaminathan, Algira, & Coulson, 1978). For personnel testing, however, the distinction is exceedingly important. The term criterion has been widely used to identify a variable external to the test itself. It is quite possible, particularly in the development of work sample tests, to construct the test so that scores on it can be directly interpreted in relation to a standard of job performance (criterion) measured externally. This may be more than simply using expectancy tables to interpret test scores, although that could be one example. It could also imply that a work sample constructed to abstract various components of the job can yield scores explicitly tied to such job performance measures as scrap rates or others. Such interpretation of scores in relation to external criteria has never been envisioned in the educational measurement literature on so-called criterion-referenced testing, but it is important enough in personnel testing to warrant special efforts to avoid confusion.

Moreover, the emphasis on content-referenced interpretation according to the Standards refers to those interpretations "where the score is directly interpreted in terms of performance at each point on the achievement continuum being measured" (APA et al., 1974, p. 19, emphasis added). This is clearly a different idea from much of the literature on criterion-referenced testing, which effectively treats any score in the distribution simply as above or below a specified score or standard.

In summary, content-referenced testing seems a preferable term because (a) it is more descriptive, (b) it avoids ambiguity, (c) it fits the terminology of the Standards, and (d) it avoids any implication of

dichotomy. The term is not the only one that might have been chosen. The relevant literature includes, in addition to content-referenced and criterion-referenced measurement, standards-referenced measurement, universe-referenced measurement, domain-referenced measurement, objective-referenced measurement, and mastery testing. Each of these terms has been proposed, and has its adherents, because of a special emphasis that is sought. This is a final advantage of the term chosen for this report, because it seems indebted to no prior bias.

The foregoing is more than a semantic exercise. The choice of language can influence substantially the directions taken in applying the diverse literature, some of which has been spawned less from an interest in making a new contribution to measurement theory than in challenging the old and established. The concept, under whatever name is chosen, has attracted very little attention among personnel testing specialists. Tenopir (1977) said that "the notion of criterion-referenced test interpretation... has no application in an employment setting" (p. 51). Ebel (1977) seems to agree. The point of their rejection of the idea may be as much a rejection of the rhetoric leading to dichotomous scoring as of the idea of interpreting scores relative to a standard.

Certainly there are places in personnel testing where one should interpret measurement against some standard other than the mean of a distribution, even if it means a dichotomous interpretation. Certainly, where productivity is determined by the speed of a moving conveyor, the individual who cannot keep up with the conveyor belt is performing at an inadequate level, whether that person is at the bottom of a distribution or merely a standard deviation below the mean.

Work Samples as Content-Referenced Tests. Work samples constitute a special form of content-referenced testing; the principal evaluation of them is in terms of job relevance. The previous discussion of content domain sampling suggested that judgments of job-relatedness can be

simplified by thinking of a four-stage process of defining the most complete possible conception of the job (the job content universe), selecting a domain of interest from that universe, and then defining the related test content universe and domain.

A work sample test is developed by sampling from that final domain. In some cases, one might use work sample techniques to develop a test which is not strictly a sample of work performance but from which work performance might be inferred. It has become an accepted cliché for such tests to refer to "the inferential leap." Figure 3 is a whimsical attempt to show graphically (and perhaps whimsically) some limits to the appropriateness of the term.

Tests can be developed which literally sample job content adding only enough testing operations to provide a scoring system. Probationary assignments can be carefully chosen, and performance on them can be carefully evaluated. These are the most complete samples that can be developed for selection or certification purposes. Simulations represent, in varying degrees, abstractions of "real" job content; they are less precisely samples, shorter, and more standardized. Tests called "work samples" are usually also abstractions from job content, typically more abstract than simulations.

The meaning of abstraction, in this context, can be illustrated by referring again to the stenographer's job. In work sample testing, one does not try to create precisely every exact task and every exact environmental condition influencing task performance. Rather, one classifies various kinds of tasks (classification is itself a process of abstraction), and creates examples of the different classes; these, performed under standard conditions and scored according to rules which are not part of the job, become the work sample test. In all three cases, the performance evaluated is a direct sample of performance on the actual job. A small problem of inference may be introduced by the scoring or evaluation

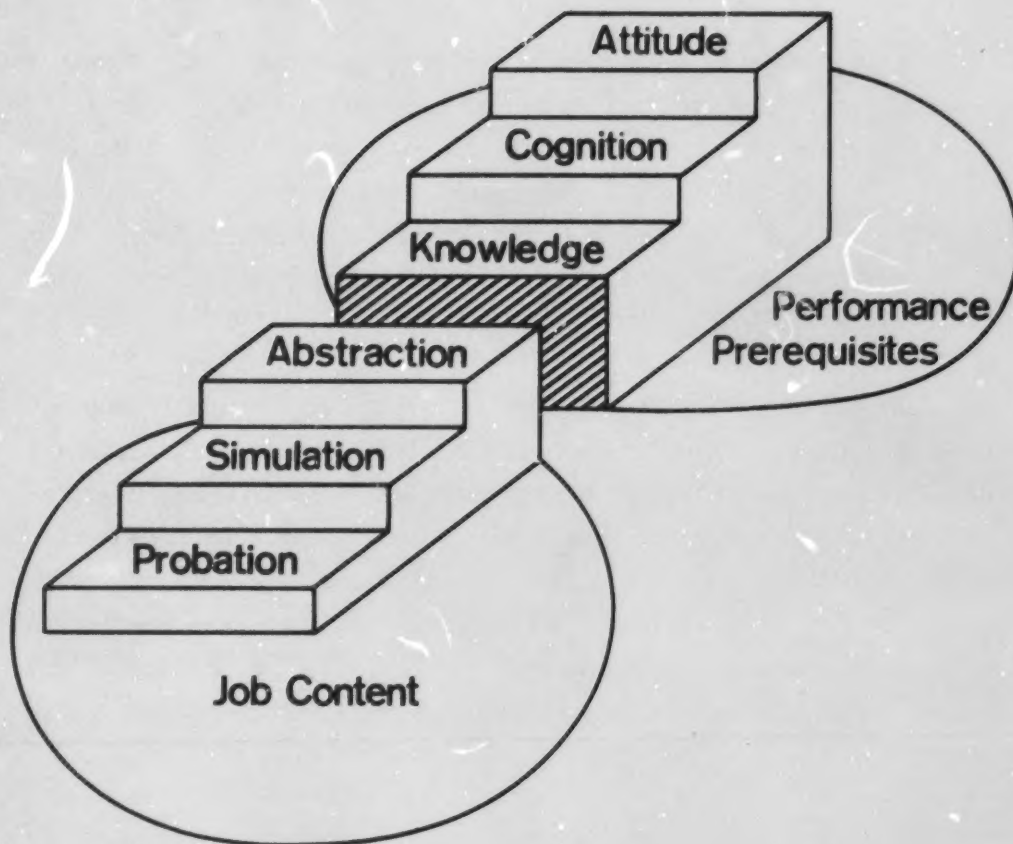


Figure 3. Samples and inferences in work sample testing

procedures, which can be contaminated by factors unrelated to real job performance, but the inference can hardly be said to require a leap.

A substantial portion of the job content domain, and therefore of an appropriate test content domain, consists of knowledge required to perform the job. In a work sample test consisting of tasks to be performed, the examinee gives evidence of the prerequisite knowledge by performing satisfactorily. In many certification programs, however, the work sample degenerates into a test of job knowledge alone. The verb has been chosen judiciously, for the job does not consist of knowledge isolated from action. (Some jobs consist primarily of knowledge. Where mastery of the knowledge component is likely to be a harder or more critical feature of the job than any actions using it, a job knowledge test is one kind of direct work sample.) The use of the job knowledge test usually implies the inference that having the knowledge leads to effective performance. Figure 3 suggests that this may not be a very great leap — more an inferential step — but that it is indeed more an inference than a sample. When one departs still further from actual performance of the job content, such as inferring prerequisite cognitive skills or essential attitudes, the measurement of these attributes really does require an inferential leap from test content to job content.

The greater the degree of abstraction from actual daily job assignments, the more appropriate is the metaphor of the leap, and also the more appropriate is a criterion-related validation strategy. Work sample testing, if it is to be accepted on its own terms as content-referenced testing, should be concerned more with sampling than with inferring.

Job Analysis. Many kinds of job analysis procedures can be used for content-domain sampling. The procedures suggested here are illustrative, not prescriptive.

Briefly, the job analysis procedure may result in a series of formula

statements of the form, " (Takes action) in (setting) when (action cue) occurs, using (tools, knowledge, or skill) ."

For a truck mechanic, such a statement might read, "Flushes truck radiator in garage when engine is said to overheat using water under pressure in flush tank." From such statements, one can specify what a worker does, what knowledge is necessary to do it, where information or material used in doing it comes from, and what happens after the task is finished. Such information defines the tasks, the methods, the prerequisites, and the contingencies that comprise the job content universe.

With the job content universe defined, panels of expert judges — people who know the job well — can whittle it down to a test content domain and can establish test specifications.

Assembling Test Content. In paper-and-pencil testing, one refers at this stage to writing items. The "items" in a work sample test might be tasks. Alternatively, tasks might be "subtests" and the "items" might be component characteristics of the process or product evaluated. In any event, scorable elements of the test are defined, developed, and assembled by experts on the job.

The essential meaning of the scores depend on the qualifications of the experts, the care with which they have reached the various judgments, and their overall degree of agreement. If all has been well done, scores (whether overall or on component tasks) can be interpreted directly with reference to the content of the test and without reference to any distribution of scores.

Scaling Test Content. Interpretation of scores with reference to test content can be facilitated and defended by establishing a formal metric for scoring. If a series of components of tasks, or components of a task content domain, can be arranged according to a genuine Guttman scale, all scores can be interpreted with reference to points on that

scale. This idea grows out of the illustration of "content standard scores" offered by Ebel (1962). In an arithmetic test of many items, ten items were selected. To this writer, merely glancing at the items identified, they seemed to fall along a scale of difficulty. If indeed they did fall along a scale, without overlapping discriminial dispersions, then any measurement technique using the other items could be tied statistically to the values along that scale. The result would be a content-referenced score with formal demonstration of transitivity.

An example of measurement approaching this sort of scaling is the Learning Assessment Program described by Grant and Bray (1970). In this program, examinees were given a series of tasks to learn to perform, seven in all; these were ordered so that it was necessary to have learned how to do task 1, to do task 2, and so on. The score for evaluating performance in this program was the level of the tasks learned. Thus, one who learned five tasks in a reasonable time was considered more proficient at the overall set of tasks than one who could only master three.

The same logic, it should be noted, can be applied to cognitive skill items. If it can be shown that a subset of items do form a reproducible scale, and if it can be further argued that these items constitute marker variables for a particular construct, then the formal properties of the scale should provide a sufficient operational definition for the evaluation of a testing program using it.

Evaluating Content-Referenced Tests. Do classical concepts of reliability and validity apply to content-referenced tests? Is it sensible to develop a test to measure, let us say, proficiency at the end of training (all trainees having at that time mastered the material of the test and therefore exhibiting no individual differences in proficiency), and to evaluate that test in classical terms defined on the basis of test score variance? Does it make sense to use norm-referenced concepts to evaluate content-referenced tests?

Much controversial literature has been devoted to such questions. The controversy probably stems from the non sequitur imbedded in the second question. It is indeed a non sequitur to equate measurement objectives with instructional objectives. A desire to have all trainees perform at an equally high level at the end of training is an objective demonstrably different from a desire to measure performance at that level. An analogy would be a Procrustean desire to stretch all little boys during their period of growth so that they can all be basketball players exactly seven feet tall. Success in the venture would lead to measures of height that have no variance; it does not follow that the yardstick used should be incapable of identifying other heights! Neither does it follow from recognizing this absurdity that variance-based statistics for determining reliability and validity are the appropriate evaluations.

In psychological measurement generally, validity has been an over-rated approach to evaluation; in work sample testing, validity concepts are far less important evaluations than are evaluations of job relevance. Content-referenced work samples developed according to the principles outlined above are assuredly job-related solely because of the method of their construction. Such a test, if scored with reference to a formal Guttman scale, could be evaluated particularly highly because of the meaningfulness of the metric. It is unfortunate that preoccupation with the concept of validity in classical measurement theory should make test users so willing to ignore the quality of measurement per se in their evaluations of the use of a test.

To assert that validity is an over-rated concept does not deny its real importance. In any sort of measurement where inferences are to be drawn beyond the descriptive character of the measuring instrument, the form of validity, generally called construct validity, is essential; nothing in content-referenced measurement relieves it of the obligation to be concerned over construct validity. Content domain sampling offers the first, and perhaps the only necessary, validity of inferences of

ability to do the job as sampled. If, however, the intended interpretation of the score seems to include something more than the test content (a frequent case), such as mastery or competence, then the score implies expectations the soundness of which must be demonstrated by the usual lines of evidence of construct validity (Linn, 1977).

That evidence may require experimental data showing that variance with groups judged as competent (or masters) is low relative to the variance between those groups and others judged as less competent (or nonmasters). Traditional validity coefficients may be useful, where obtained variances permit them, as results of inquiries into different aspects of the construct validity of the scores. Also, scores (or observations) on content-referenced tests must be reliably determined, although the nature of reliability may be conventional estimates of systematic variance, studies of the generalizability of scores, or the consistencies of classifications.

LATENT TRAIT THEORY

Under various names (latent structure analysis, item characteristic curve theory, Rasch model), latent trait theories constitute another approach to the construction of formal measuring instruments. The distinguishing importance of the method is that it defines item difficulties and other characteristics more or less independently of characteristics of the particular samples from which the data distributions are drawn.

Originally developed for the assessment of attitudes (Lazarsfeld, 1950), latent trait theory has subsequently been used mainly in the measurement of cognitive abilities (Lord & Novick, 1968; Hambleton & Cook, 1977). It can be used for at least some forms of work sample testing. Applications to tests of knowledge have been shown by Bejar, Weiss, & Gialluca (1977), and an application to personality measurement by Bejar (1977) seems directly applicable to measures of the quality of work sample products and other practical problems of personnel testing.

The Theoretical Foundations. Although the mathematical foundations of latent trait theory are beyond both the scope of this report and the abilities of the writer, a brief account of the nature of the theory is useful for discussions of its applicability.

An item characteristic curve can be identified in which the probability of a correct response to the item is seen as a function of the examinee's ability level. Various models exist for defining the function, one of which describes the item characteristic curve as a normal ogive. Figure 4 shows hypothetical item characteristic curves for three items. Item 1 has a fairly typical difficulty level; many people get it wrong, but many get it right. Item 2 is a very difficult item; only people of very high ability are likely to get it right, although people of low ability seem to get it right by guessing more often than on the other items. Item 3 is a highly discriminating item; most people with above average ability will get it right, and those with ability below average are unlikely to give a correct response.

Three parameters can be estimated for defining each of these curves. Parameter a is a discrimination index, proportional to the slope of the curve at the inflection point. Parameter b is a difficulty index, defined as the ability level on the base line corresponding to the point of inflection (the point corresponding to a .50 probability of correct response if the third parameter is zero). Parameter c is the probability of a correct response at infinitely low ability levels, often called the guessing parameter. Parameters estimated in a given analysis include the ability levels, identified as theta (θ) in Figure 4, of the people tested as well as the item parameters.

The theta scale may be defined arbitrarily in any given analysis; the numerical values of the difficulty parameters are therefore arbitrarily expressed for a given sample. However, parameters estimated from samples with different characteristics correlate very highly, even if one

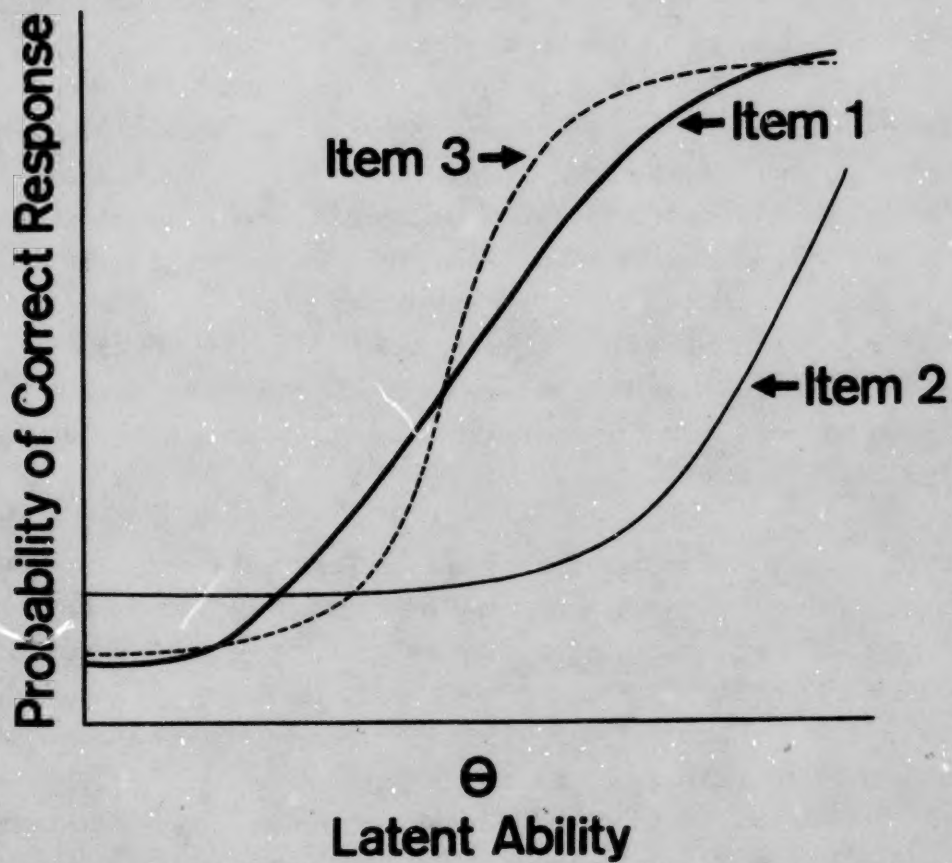


Figure 4. Item characteristic curves of three hypothetical items

sample consists of the low-scoring half of a distribution and the other sample consists of the high-scoring half (Rudner, 1976). Available equating procedures permit merging the latent ability scales for the population as a whole and expressing item characteristics in terms of that common scale. The resulting item characteristic curves are essentially congruent regardless of the sample from which they were developed. Failure to obtain such congruence indicates either a poor fit of the model or the possibility of item bias (Ironson, 1977).

The description presented here with Figure 1 (three parameters defining a normal ogive) refers to one of many models for latent trait analysis. There are logistic curves as well as normal ogives, and there are models that estimate only one or two of the parameters. The "two-parameter" models estimate discrimination values and difficulties; the "one-parameter" models estimate only difficulty levels. Multidimensional as well as unidimensional models have been proposed, and models are available for dichotomous, polychotomous, graded, or continuous responses (Samejima, 1969, 1972, 1973).

In classical psychometric theory, the standard error of measurement is generally treated as equal across the range of a distribution of scores. Its counterpart in latent trait analysis, the standard error of the estimate of ability, varies with the ability level. It is possible to construct item information curves showing the precision of the estimation of ability from responses on a single item at different ability levels. Tests measuring the same latent ability on the common scale can be assembled with different combinations of items, each with different item characteristic curves and item information curves. Combining item information curves across items yields a test information curve, the highest point of which is the level of ability at which the information (that is, the precision of estimating ability) provided by that set of items is greatest. Item characteristic curves may likewise be combined to yield a test characteristic curve in which the probability of an obtained

score is a function of the underlying ability level.

Uses of Latent Trait Analysis. If, for a given item, the item characteristic curves for two distinguishable groups of people are not essentially congruent, then that item cannot be said to be measuring the same latent ability in those two groups. Therefore, latent trait theory can be used to identify sources of item bias across race or sex groups.

This has implications for judgments about the adverse impact of tests used as decision tools. If there are substantial differences in obtained score distributions, the proportions of the groups selected (or classified into a desirable category) will differ. Current governmental regulations governing the use of employment procedures call for investigations to determine which of alternative selection tools will have lesser adverse effect, that is, which tests will have smaller mean differences in test performance.

If there are true subgroup differences, psychometric properties of the tests may affect the size of the adverse effect. Highly unreliable tests will have little adverse effect, for example. The problem can be highlighted by looking at test characteristic curves. The true differences in ability (as shown by the mean estimates of latent ability) are not influenced by the choice of test, but observed differences are. A test with a smaller slope on that curve will show less adverse effect than will a test with a characteristic curve that is steeper. In other words, even though the true differences are not changed by changing the test, the observed differences may be markedly greater for one test than for another — and both can err in opposite directions. One of the tests may falsely exaggerate the true differences, while the other may falsely minimize them.

Working from item and test information curves, one can assemble small sets of items yielding the most precise possible ability estimates

at different ranges of ability levels (Lord, 1968; Weiss, 1974). If care has been taken to assure a full-range scale of ability in the development of an item bank, with known item characteristic curves, then any individual can be tested and located along that scale even using a unique set of items. Once the individual is located on that scale, the interpretation of his score is content referenced. For personnel testing, tests can be tailored not only for individuals but also for individual jobs requiring different levels of a particular ability, and standards for each job can be defined in terms of the ability levels appropriate.

Evaluation. Tests constructed using latent trait analysis can be evaluated with conventional concerns for job relatedness, reliability, and validity, but they may be better evaluated in other ways.

Job relatedness of work samples constructed by latent trait studies is no different from job relatedness of other work sample. In either case, it depends on the quality of the judgments made in defining the job content universe and in moving logically from that definition to a set of test specifications. Latent trait theory may, however, make it possible to develop abbreviated work samples that will be equally job related by identifying components that will maximize information at different levels of proficiency.

In latent trait theory, classical reliability is replaced by the idea of the information curve. Reliability coefficients can be manipulated by manipulating samples (Samejima, 1977); they are not sample-free. The standard error of measurement is a general statistic applying to all examinees in a distribution (or, if specially computed, in a specified broad range of the distribution). The standard error of estimate, however, is a value describing the precision of measurement at a particular point on the ability scale and is therefore far more informative. The test information curve gives evaluative information similar to that provided by reliability coefficients, but it does it better.

Construct validity is less important in latent trait studies than the fit of data to a model. If the data obtained from the items will indeed fit a latent trait model, they are certainly measuring something and doing so with internal consistency. Item construction proceeds, of course, in the context of a particular construct, so it is not difficult to define the underlying trait dimension. Construct validity, if of interest, is further assured if biased items (or items with other evidence of poor fit) are eliminated from the test or item pool as potential sources of contamination.

In general, however, validity statements are superfluous. The amount of research that goes into the development of such tests is indeed substantial. When that research has been completed, and measurement is expressed in terms of the underlying scale, that measurement is a sufficiently satisfactory operational definition of the construct being measured; no additional recourse to concepts of validity is necessary or informative.

GENERALIZABILITY THEORY

Generalizability theory (Cronbach, et al., 1972) does not challenge the norm-referenced basis of classical psychometric theory; it is, in fact, an extension of classical theory. The challenge it poses is the challenge to the undifferentiated distribution of error implicit in the classical formulation of true scores and error scores comprising an obtained score. Moreover, estimation of error in psychometric theory is built on the requirement of parallel tests, a condition not regularly satisfied in psychological measurement.

Any observed score is based on measurement obtained under a specified set of conditions. That set of conditions is but a sample of all of the possible sets that might have existed. Recognizing this, Cronbach and his associates ask investigators to define the universe of conditions, or the universe of possible observations, under which a person might be

tested. One generalizes in any actual use of tests from the sample to a universe of applicable conditions; generalizability studies make it possible to define the limits of possible generalizability for any test, a result particularly valuable in work sample testing.

An illustration of this implication may be helpful. Suppose that a work sample test is devised for measuring a specified skill at the end of training. Suppose, moreover, that the test is administered under traditional ideas of good test administration: good lighting, giving instructions carefully and consistently, special efforts to ascertain the reliability of observation, and a general effort to provide conditions optimally suited for maximizing performance of the examinee or reliability of the observations.

Now, no one is really interested specifically in how well the individual performs at the end of training except possibly the trainers. From an organizational point of view, the measurement of skill at the end of training is intended to generalize to conditions less optimal but more realistic, that is, to field rather than institutional settings. Obviously, there can be many different kinds of field conditions. Conditions can vary according to light sources, according to geographical climate, or according to variations in degrees of situational hostility.

A generalizability study, or a multiple facet analysis as it is also known, can be designed to determine the degree to which scores obtained in a sample measured under optimal conditions can be generalized to the different, non-optimal conditions of the study. Three possible kinds of findings can emerge: one may find that the inferences generalize quite well across conditions, one may find that they generalize not at all, or one may find that they will generalize to a limited subset of conditions, that is, that generalization across facets is possible only by the deletion of certain conditions.

One other point, too important for the possible implications of generalizability theory for work sample testing to be omitted from this brief discussion, is that the method permits one to estimate universe scores or expected obtained scores under specifiable combinations of facets. That is, even if there are substantial differences in performance under different sets of conditions, one may be able to generalize beyond the initial condition by making estimates of the obtained scores that would be expected under specified kinds of field conditions.

Program Evaluation. Alternatives to conventional validation procedures include evaluations of total programs using personnel tests. The use of assessment centers, in particular, has led to a situation in which the predictor is no longer a single test or small battery but the outcome of a complex assessment procedure expressed as the judgment of observers.

A less formal version of the same kind of thing occurs in an employment office where, instead of using a test and expectancy chart or cutting score, a series of assessment devices will be selected depending on the questions a decision-maker wishes to answer about a particular candidate for a particular job. Different batteries of tests may be used, different weights may be given to the same tests, and different questions may be asked. The procedure is frequently called clinical judgment or clinical prediction.

The total testing program, including judgments or decisions, can be evaluated in such circumstances if a quasi-experimental design can be used to compare the effectiveness of the performance, or work force stability, or other outcome in organization using the program to that in a different organization, reasonably well matched with the first, in which the program is not in use.

In a sense, this is criterion-related validation of the final judgment. It is, however, more in line with modern concerns for program

evaluation, and it is mentioned here as a potential stimulus to exploring the literature on program evaluation for its possible implications in the evaluation of personnel testing programs.

SUMMARY

Personnel testing programs have traditionally been evaluated in terms of the classical psychometric concepts of validity, particularly of criterion-related validity. The habit is well entrenched. Both the Standards (APA, et al., 1974) and the Principles for the Validation and Use of Personnel Selection Procedures (Division of Industrial-Organizational Psychology, 1975) give institutional support and encouragement to the habit. It is not a bad habit, like smoking, hazardous to the user's health and therefore to be broken; rather it is like eating, a habit to be tempered with moderation. Classical notions of validity have been valuable, but there are evaluative concepts that are more useful for some uses.

One of the difficulties with classical notions of validity is that there are too many of them and, in personnel testing, they have been forced to fit into too many Procrustean beds. The basic notion of validity as an evaluation of measurement has been stretched into something called content validity and squeezed into something else called criterion-related validity, neither of which refers to the quality or meaningfulness of measurement per se. Only investigations of construct validity provide useful insights into the meaning of measurement; what is called content validity is really better understood as content-oriented test development, and criterion-related validity is in reality the outcome of a test of a hypothesis.

In personnel testing, criterion-related validity holds a place of high honor. It is an established, useful approach for demonstrating the relationship of performance on the test to performance on the job --

a phrase which, when abbreviated, becomes job relatedness.

Job relatedness, or job relevance, is the most important single consideration in the evaluation of most personnel testing procedures, whether the testing is used to predict future performance, certify competency, evaluate performance, or validate some other variable. Criterion-related validity is a good source of evidence for judging the job relatedness of a test, but it is not the only one.

Equally important evidence of job relatedness is showing that the test is an acceptable operational definition of important aspects of job performance. Such a showing is based primarily on a thorough, rational process of getting information about a job and using expert opinion in defining domains, test specifications, and the relevance of individual items within the test. Surely, such information is at least on par with evidence of criterion-related validity serendipitously found using a test developed for a wide variety of general uses.

Another vitally important consideration in the evaluation of a test is the meaningfulness of scores obtained through its use. Meaningfulness can be established in part through the methods of establishing construct validity or from the methods of test construction. A very specific kind of meaning is derived through criterion-related studies. A quite different but perhaps equally valuable source of meaning is the concept of a latent trait.

In short, a score on a personnel test becomes meaningful in a variety of ways. It is meaningful if it can be interpreted in terms of a predicted level of future performance or of a probability of attaining some stated level of performance. It is meaningful if it can be interpreted as a proficiency measure on a sample of the actual job. It is meaningful if it can be interpreted directly in terms of a standard performance or in terms of a scale reflecting the variable being measured

without reference to an idiosyncratic distribution obtained from an available sample of people — or, for that matter, of items. Its meaningfulness is enhanced to whatever degree it can be expressed as a score on a meaningful scale which retains that meaningfulness over a wide variety of circumstances. A content-referenced interpretation is at least as meaningful as a criterion-referenced interpretation (using the term here in its unusual sense of a score interpreted in terms of an external criterion variable). Thus methods of scaling or calibrating tests (such as latent trait analysis) need to be given a priority at least as high as that given to criterion-related validation in evaluating the meaningfulness of scores.

Classical test theory also evaluates tests in terms of reliability, meaning the freedom within a distribution of test scores from variance due to random error. Classical notions of reliability do not take systematic error into account. The application of the reliability concept to the evaluation of a single score is through the standard error of measurement, a value generally taken to be the same throughout the entire distribution of scores.

These are also useful evaluations, but they, too, can be improved upon through the use of newer ideas. Latent trait theory, for example, replaces the reliability theme with the idea of the information curve, using the standard error of estimated abilities as an index of precision at specific ability levels. Generalizability theory offers a much more comprehensive and useful accounting of various sources of error and their magnitudes, and it permits statements of both the limits of generalizability and the estimates of scores in different sets of conditions.

Modern measurement theory, although it has offered challenges to classical psychometric theory, has not reduced the usefulness of classical evaluations, especially in situations such as the use of tests or ratings in the measurement of such variables as attitude or personality

characteristics. For many other variables and for other methods of measurement, however, personnel testing needs to explore and exploit the possibilities of the newer theories. These possibilities are particularly relevant to work sample testing because it is most appropriately evaluated in terms of job relevance and its amenability to content-referenced interpretations of scores.

REFERENCES

- American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.
- Bejar, I. I. An application of the continuous response level model to personality measurement. Applied Psychological Measurement, 1977, 1, 509-521.
- Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of convocational and adaptive tests in the measurement of classroom achievement. (Resch. Rep. 77-7). Minneapolis: University of Minnesota, Psychometric Methods Program, 1977.
- Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) Washington, D.C.: American Council on Education, 1971.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurement. New York: Wiley, 1972.
- Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.
- Division of Industrial-Organizational Psychology. Principles for the validation and use of personnel selection procedures. Dayton: Author, 1975.
- Ebel, R. L. Obtaining and reporting evidence on content validity. Educational and Psychological Measurement, 1956, 16, 269-282.
- Ebel, R. L. Must all tests be valid? American Psychologist, 1961, 16, 640-647.
- Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.
- Ebel, R. L. Comments on some problems of employment testing. Personnel Psychology, 1977, 30, 55-63.
- Grant, D. L., & Bray, D. W. Validation of employment tests for telephone company installation and repair occupations. Journal of Applied Psychology, 1970, 54, 7-14.
- Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

- Guion, R. M. Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.
- Guion, R. M. Content validity — the source of my discontent. Applied Psychological Measurement, 1977, 1, 1-10.
- Guion, R. M. Scoring of content domain samples: The problem of fairness. Journal of Applied Psychology, in press.
- Gulliksen, H. Intrinsic validity. American Psychologist, 1950, 5, 511-517.
- Hambleton, R. K., & Cock, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.
- Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.
- Hull, C. L. Aptitude testing. Yonkers, N.Y.: Work Book, 1928.
- Ironson, G. H. A comparative study of several methods of assessing item bias. Unpublished doctoral dissertation, University of Wisconsin-Madison, 1977.
- Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. Measurement and prediction. New York: Wiley, 1950.
- Linn, R. L. Issues of validity in measurement for competency-based programs. Paper presented at the meeting of the National Council of Measurement in Education, New York, April, 1977.
- Lord, F. M. Some test theory for tailor testing (ETS RB-68-38). Princeton, N.J.: Educational Testing Service, 1968.
- Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.
- Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.
- Rudner, L. M. Item and format bias and appropriateness. Washington, D.C.: Model Secondary School for the Deaf, 1976.

- Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph No. 17, 1969.
- Samejima, F. A general model for free-response data. Psychometrika Monograph No. 18, 1972.
- Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38, 203-219.
- Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247.
- Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validation studies. Journal of Applied Psychology, 1976, 61, 473-485.
- Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.), Washington, D.C.: American Council on Education, 1971.
- Tenopyr, M. L. Content-construct confusion. Personnel Psychology, 1977, 30, 47-54.
- Weiss, D. J. Strategies of adaptive ability measurement (Resch. Rep. 74-5). Minneapolis: University of Minnesota, Psychometric Methods Program, 1974.
- Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on testing problems. Princeton, N.J.: Educational Testing Service, 1968.



ARI TECHNICAL REPORT
TR-79-A10

Principles of Work Sample Testing: III. Construction and Evaluation of Work Sample Tests

by

Robert M. Guion

BOWLING GREEN STATE UNIVERSITY
Bowling Green, Ohio 43403

April 1979

Contract DAHC 19-77-C-0007

Prepared for



U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Approved for public release; distribution unlimited.

800555 3/4

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR-79-A10	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PRINCIPLES OF WORK SAMPLE TESTING: III. CONSTRUCTION AND EVALUATION OF WORK SAMPLE TESTS		5. TYPE OF REPORT & PERIOD COVERED Final 15 Nov 1976 - 15 June 1978
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert M. Guion		8. CONTRACT OR GRANT NUMBER(s) DAHC19-77-C-0007
9. PERFORMING ORGANIZATION NAME AND ADDRESS Bowling Green State University Bowling Green, Ohio 43403		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q161102B74F
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, Virginia 22333		12. REPORT DATE April 1979
		13. NUMBER OF PAGES 89
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) --		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) --		
18. SUPPLEMENTARY NOTES Monitored by G. Gary Boycan, Engagement Simulation Technical Area, Army Research Institute.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Measurement theory, psychometrics, work sample testing, validity, content- referenced testing, criterion-referenced testing, latent trait theory, generalizability theory		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Work sample tests should be relevant to the job, objectively constructed and scored, reliable, and capable of being scored on a standardized content- referenced scale. Detailed steps in working from job analysis to establishing test specifications are presented for assuring job relevance. Methods are suggested for developing scales for scoring by a priori scaling, or by latent trait analysis, to provide a standard, content-referenced scale for scoring. Job samples should be evaluated primarily in terms of relevance and of generalizability. Seven principles of work sample testing are offered to		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE

AUG 25 1980

133

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

20. (continued)

researchers.

This report, the third of four, is written for psychologists and others interested in research testing.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

PRINCIPLES OF WORK SAMPLE TESTING: III. CONSTRUCTION AND EVALUATION OF WORK SAMPLE TESTS

BRIEF

Desirable characteristics of work sample tests include standard scales for content-referenced scoring, adequate variance for statistical analysis, objectivity in structure and scoring, reliability of both measures and classifications, functional unity in measurement, and job relevance.

To assure job relevance, a rather deliberate process of domain definition is recommended. It begins with a thorough definition of the total job, a job content universe, defined in terms of component tasks, checklist-defined task elements, and clusters of similar component tasks called task categories. From the total universe, a sample (non-random) is selected defining the portion of the universe important for testing; this portion is the job content domain. All possible ways test tasks or items could be constructed and scored to sample performance in the job content domain define a test content universe, a universe of all possible admissible operations. Selecting from this universe, again with no attempt to be random or representative, is based on practical considerations of time, cost, and measurement feasibility and yields a portion of the possible universe which is the test content domain. From this domain, test specifications are prepared and, if the test is developed according to those specifications, either (a) it will be unarguably job relevant, or (b) the question of relevance will be directed toward the domain definitions rather than toward the test.

With test specifications established, a work sample test is developed by a panel of qualified experts who are knowledgeable about the job and who have also had some training in the preparation of test items. Items for work sample tests may include conventional written items testing job knowledge or practical items based on the total job, on some direct sample of the job domain, or an abstraction from the domain. Methods are presented for systematizing panel judgments of item relevance.

Scores on the test should, for a work sample, be interpretable in terms of the content rather than only with reference to norm groups. That is, work sample tests should be content-referenced rather than norm-referenced. Some alternative methods of scoring are presented. An emphasis is placed on test development and scoring using latent trait theory, which provides a standard content-referenced interpretation that can be independent of the distribution of a given sample

of people and of the particular set of items chosen for testing a given individual. An alternative system involving a priori scaling is also discussed. The practice of reducing all scores to a pass-fail dichotomy is discussed (procedures for doing so are briefly presented) and discouraged. For work sample tests, level of mastery seems more important from a measurement perspective than classification as a master; degree and direction of misclassification is more important than the mere fact of misclassification.

Generalizability theory is briefly discussed in this report, although a more detailed discussion of it is reserved for the fourth report of the series.

The evaluation of a work sample test should emphasize job relevance and generalizability. Other evaluations to be considered include reliability (with perhaps greater emphasis on the reliability of the scores as measures than on the reliability of classifications), accuracy of measurement, information curves, construct validity (with an emphasis on examining alternative explanations for scores), and in some cases evidence of predictive or other criterion-related correlations.

The paper concludes with seven basic principles of work sample testing: (1) choices of job content domains need justification, (2) test content domains should be as congruent as possible, (3) scoring procedures should strive toward fundamental measurement, emphasizing transitivity within a reasonably homogeneous domain, (4) scores should permit assessment of levels of proficiency rather than mere dichotomies, (5) opportunities for irrelevant influences on individual scores should be minimized, (6) scoring of work sample tests designed for use in large, multi-location organizations, should be standardized on a content-referenced scale applicable to the organization as a whole, and (7) scores on a work sample test given in a setting of institutional control and standardization should generalize to a variety of field settings.

TABLE OF CONTENTS

INTRODUCTION.	1
EXAMPLES OF WORK SAMPLE TESTING.	2
DESIRABLE CHARACTERISTICS OF WORK SAMPLES	6
STANDARD CONTENT-REFERENCED SCORES.	7
VARIANCE	9
PSYCHOMETRIC CONSIDERATIONS	10
PREVIEW.	12
DOMAIN DEFINITIONS.	13
DEFINITION OF JOB CONTENT UNIVERSE.	15
DEFINITION OF JOB CONTENT DOMAIN	18
DEFINITION OF TEST CONTENT UNIVERSE	20
The Universe of Admissible Observations	21
Item Forms	22
DEFINITION OF TEST CONTENT DOMAIN	23
Test Specifications	24
TEST DEVELOPMENT	25
QUALIFICATIONS OF ITEM DEVELOPERS	25
ITEM POOLS.	28
Job Knowledge Items	28
Items in the Total Job Test	30
Items in Direct Work Samples	31
Items in Abstracted Work Samples.	33
Requirements for Scorable Items	36
ITEM ANALYSIS.	36
Retranslations.	37
Content Validity Ratio	37
Index of Item-Objective Congruence	39
Conventional Item Analysis.	41
SCORING AND CALIBRATION	43

TABLE OF CONTENTS

LATENT TRAIT ANALYSIS.	44
Other Latent Trait Models.	50
Scoring Procedures	51
Advantages of Latent Trait Analysis	54
Possible Alternatives to Latent Trait Theory	58
CUTTING SCORES	60
GENERALIZABILITY ANALYSIS	64
EVALUATIONS OF WORK SAMPLES	68
JOB RELATEDNESS.	68
RELIABILITY	69
Reliability of Mastery Classification.	69
Reliability of Measures	70
A Coefficient of Accuracy.	74
Modern Replacements for Reliability	74
VALIDITY	76
Construct Validity	76
Consistency in Domain Sampling	77
Predictive Utility	78
SUMMARY: PRINCIPLES OF WORK SAMPLE TESTING	79
REFERENCES	84

LIST OF FIGURES

<u>Figure No.</u>		<u>Page</u>
1	Three-Parameter Item Characteristic Curves, for Four Hypothetical Items, Showing Hang- ing Order of Difficulty at Different Ability Levels.	46
2	Obtained Score Differences Between Groups as a Function of Different Test Character- istic Curve Parameters.	57
3	Optional Cutting Score as the Point of Interaction of Score Distributions of Masters and of Non-Masters.	63
4	Venn Diagram Identifying Variance Estimates for Persons (p) Nested in Observers (o) Nested in Installations (i), all Crossed with Conditions (c).	65

INTRODUCTION

At least since World War I (Yerkes, 1921), paper-and-pencil testing has been accepted as the prototype for psychometric theory. The massive testing programs initiated and carried out by the Army during that war were expanded in military services and in employment settings during the years between wars. By World War II, psychological testing for classification purposes, primarily using paper-and-pencil tests, was widespread. The Aviation Psychological Program (Guilford, 1947) used a few apparatus tests, such as measures of coordination and dexterity, but these were oddities in the general classification test batteries. Since World War II, testing practice has continued to be dominated by paper-and-pencil tests, mainly measures of knowledge or of cognitive variables.

Concomitantly with the growth of the use of the paper-and-pencil tests was growth in the development of the psychometric theory. Beginning with Thurstone's famous monograph on the primary mental abilities (Thurstone, 1938), through the factor analytic work in the Aviation Psychology Program, the publication of the monumental textbook by Lord & Novick (1968), and a substantial literature since, the theory of mental tests scores has moved from a few brief equations into an astonishingly complex mathematical structure.

No corresponding effort has been expended in the measurement of performance variables. Performance testing also dates back to before World War I, but in the Army, these tests were limited pretty much to testing for intelligence, not proficiency at designated tasks. The Stenquist Test of Mechanical Proficiency (requiring the assembly of common objects) was used as a group test of intelligence for "illiterates and foreigners." It was replaced by the beta examination.

"The chief objection to it was its low value as a measure of intelligence. Even with unselected literate men it correlated with examination a only to the extent of 0.45 to 0.55" (Yerkes, 1921, p. 321). If such tests had been evaluated as measures of performance variables rather than as measures of cognitive variables, and if the level of research and theory given to performance measurement even partially matched the work done on cognitive measurement, the potential value of measurement by work samples might have been realized long ago. Work samples yield more nearly fundamental measurement with less reliance on norms and with potentially less likelihood of contaminating sources of variance. In contemporary society, a major problem with paper-and-pencil testing is the charge of bias. Since bias seems largely due to irrelevant variance, performance variables measured directly by performance techniques may either show less evidence of discriminatory impact against women and minorities or be taken as evidence that observed group differences are real.

EXAMPLES OF WORK SAMPLE TESTING

A work sample test is defined here as any standardized and scorable procedure in which people are asked to answer questions, solve problems, produce or modify objects, or otherwise demonstrate knowledge and competence in tasks drawn from a job content domain. This is a broad definition. It is broad enough to include as a work sample test a systematic set of probationary assignments on which performance is systematically evaluated or "scored." It can also include, as a different extreme, a paper-and-pencil test of the knowledge identified as part of a job content domain. It can include literal job assignments, or simulations that faithfully reproduce aspects of such assignments, or abstractions from job assignments that appear to be artificial but reproduce essential or crucial aspects of an assignment.

It includes tests developed as criterion measures, training devices, predictors, or standards for certification. However, it does not include tests that involve varieties of common tasks, even manipulative tasks, unless performance on those tasks can be shown to be part of a specified job content domain. By this definition, a typical test of typing speed and accuracy, for example, does not become a work sample test until the job content domain it samples is specified.

The variety of possible kinds of work sample proficiency tests is limited only by the variety of jobs and the imaginations of the test developers. Asher and Sciarrino (1974) cited over eighty published accounts of work sample tests -- with virtually no overlapping examinations.

A six-year study of bias conducted by the Educational Testing Service in cooperation with the United States Civil Service Commission (Campbell, Crooks, Mahoney, & Rock, 1973) used work sample criteria for two of the three occupations studied. An attempt to develop a work sample for medical technicians was abandoned as unsuccessful. However, three kinds of work samples were successfully developed for Cartographic Technicians: one involving the compilation of contour lines, another extracting drainage system and cultural detail from vertical aerial photographs, and the third requiring a geometric restitution of information from oblique aerial photographs. An In-basket exercise was developed for Inventory Managers; it simulated decisions and communications concerning inventory following the Military Standard Requisitioning and Issue Procedure (MILSTRIP). A hypothetical new agency was "created" for the exercise.

Some work samples involve a great deal of ingenuity in their construction. Rubinsky and Smith (1973) simulated a bench grinder's

job for an experiment in safety training. The simulation was so organized that an "accident" was signaled by turning on water jets. If the operator was standing in the correct place, he was not sprayed ("injured"); otherwise, he got wet.

In what ranks high as an understatement, Root, Epstein, Steinhiser, Hayes, Wood, Sulzen, Burgess, Mirabella, Erwin, and Johnson (1976) described the background for REALTRAIN in these terms, "A combat environment, which involves the violent interaction of two mobile opposing forces who are out to destroy one another, is difficult to simulate" (p. 1). The REALTRAIN exercise, in its various stages of development, seems to have had applied a substantial degree of ingenuity in simulating the critical component of combat experience, the knowledge of whether one has killed or been killed. In the REALTRAIN exercise, an initial procedure used telescopic sights for identifying a number on the helmet of an opposing soldier. Upon reading the other soldier's number through the telescopic sight, the soldier fired a blank round and reported the number to a controller who, by radio, identified the casualty to a controller with the opposing force. Knowledge of results with this technique was accomplished in five to ten seconds (Schriver, Mathers, Griffin, Jones, Word, Root, & Hayes, 1975). In describing the REALTRAIN exercise, Uhlaner, Drucker, and Camm (1977) reported an adaptation using lasers to simulate the direct fire characteristics of a number of weapons. Firing a blank round keys the firing of the laser; the method reduces the discrepancy between accurately sighting a target and accurately hitting it.

An abstract work sample reported by Grant and Bray (1970) was developed originally as a training device called the Learning Assessment Program (LAP). The LAP abstracts skilled activities from seven levels of telephone installation craft work. The first four refer to

fairly conventional kinds of telephone installation, while the last three are abstracted from higher level jobs. A person tested under the Learning Assessment Program could have up to three weeks to learn and demonstrate the seven levels of proficiency; scores were highest level completed, highest level passed, and time taken to complete the program.

In some jobs a simulation exercise can be required and organized so that the thought processes and judgments can be tested using multiple-choice test items. An example is the examination procedure of the National Council of Architectural Registration Boards (NCARB, 1976). In its 1975 examination, the task was to design a performing arts center for Scottsdale, Arizona. The four-part, two-day examination included detailed information about the community, the needs, legal requirements, and other matters. A variety of environmental, programming, design, and construction questions were asked. The examination materials, including actual items, was subsequently published (NCARB, 1976) along with the announcement that the 1976 examination would center on designing a facility for a prison infirmary and health center. Approximately eight pages of advance information about the examination were presented along with a substantial bibliography of useful background books and articles.

An architect's work is essentially information processing; the quality of his final product is a matter of taste, but the necessary thought processes leading to that product are objectively known; a paper-and-pencil test can therefore be a satisfactory work sample. In some jobs, however, excellence in the work process is rather irrelevant if the outcome is poor. Also, some jobs are so designed that an individual cannot do them alone; the measurement of proficiency must use a work group rather than an individual as the unit of analysis.

An example of both problems is tank warfare. An excellent knowledge of battle theory or weapon nomenclature is of little value if the tank crew cannot hit a target and makes itself vulnerable to being hit; moreover, hitting targets and staying protected is not a one-man job. A proficiency test of tank crew gunnery has been developed and reported by Wheaton, Fingerman, and Boycan (1978). It consists of a set of simulated test engagements involving different types of targets, different required behaviors of individual crew members, and some practical constraints on the use of main gun ammunition.

Work sample tests, with an emphasis on job knowledge, are used in many state or trade licensing or certification programs. They vary not only in content but in quality. Shimberg, Esser, and Kruger (1973) have reviewed licensing practices and policies in different states and in different occupations, and they present a dismaying picture. Most such tests are written examinations; Shimberg, et al. identify problems with written examinations under four headings: lack of planning, over-reliance on unreliably scored essay tests, poor quality where multiple-choice questions are devised, and lack of item analysis. Even the performance portions of such tests are frequently inadequate because of failure to sample crucial skills adequately, the failure to standardize procedures, and the lack of reliable or appropriate scoring procedures for evaluating performance.

DESIRABLE CHARACTERISTICS OF WORK SAMPLES

A work sample test is an operational definition of proficiency in some aspect of performance of the work sampled. The requirements for evaluating a test simply as a satisfactory operational definition of proficiency level, as outlined in the second report in this series, are always important in work sample testing. If these have been

satisfied, it seems unnecessary in most cases to consider still further requirements.

In many certification testing situations, however, some further problems intrude themselves. These situations include state or trade association certification examinations, or qualifying examinations for promotions, where the examinations are to be administered repeatedly or at various times and places in a multi-location organization. The Army skill qualification testing program is an example of the latter. The organization is by no means small, the jobs occur in many locations throughout the world, and the necessity to certify qualifications is a continual one. In these circumstances, there are some additional questions to be considered in the construction and evaluation of work sample tests.

STANDARD CONTENT-REFERENCED SCORES

Scores on a work sample should be interpretable in terms of test content and its relationship and the interpretation should be standard across examiners, locations, times, and conditions. There are several ways to standardize interpretations of scores. The simplest, although it is the most difficult to defend, is to interpret each score as either above or below an arbitrary cutoff point classifying examinees either as masters or as non-masters. Much of the literature on criterion-referenced testing in education seems bent on such a loss of information, and loss of information is the best description of most dichotomous scoring. It is much more useful to refer to the level or degree of mastery, a polychotomous scoring system.

When we think of standard scores, we typically think of the z-score scale, with its mean of zero and a standard deviation of 1,

or of a linear or nonlinear transformation of it. Many linear transformations have been used; nonlinear transformations are ordinarily chosen in an effort to normalize obtained distributions. Examples include the familiar stanine or McCall's T-score with its mean of 50 and standard deviation of 10. These are the standard norm-referenced interpretations of scores, and they can be quite useful for many purposes. They are, for example, extremely useful for interpreting scores on aptitude tests or on measures of personality or attitude variables.

They are not, however, desirable interpretations of scores for work sample performance. A T-score between 40 and 60, for example, means only that the examinee performs about like nearly everyone else, at least in the normative sample, but it tells nothing about his level of mastery. Some form of content-referenced scale is usually necessary to provide adequate meaning for a work sample test. At least three kinds of content-referenced scales can be devised.

1. Occasionally, a group of expert judges, considering the examination in detail, will arrive at a system for establishing an arbitrary cutting point or standard above which mastery may be claimed. Where such a standard is established, scores can be interpreted in terms of linear distances from that standard point. This can be a useful scale, but its value depends on how widely the standard is accepted.
2. A priori scaling can provide a basic reference scale. If a subset of test components or items form a scale, then total scores can be interpreted with reference to that scale of selected items.
3. If latent trait analysis is used, the test can be scored on the basis of maximum likelihood estimates or other estimates along a "sample-free" scale of underlying latent ability.

There is a special advantage associated with the last two examples.

If a work sample can be scored on an absolute scale, whether this scale be calibrated by latent trait analysis or by older scaling techniques, it will solve a long-standing problem in criterion-related validation. If the criterion is measured on such a scale, then a predictor variable that has been used to select people can be revalidated despite sample homogeneity. As Peterson and Wallace (1966) pointed out, the use of a valid predictor often results in such restricted variance that evidence of validity can no longer be obtained by computing validity coefficients. If the criterion measure is a work sample, however, and can be predicted with absolute predictions, then follow-up validation is possible by computing the variance of the errors in prediction and comparing it to the overall criterion variance in the original validation study.

VARIANCE

There has been a substantial controversy over the importance of variance in work sample testing (Popham & Husek, 1969; Millman & Popham, 1974; Woodson, 1974). The issue seems to center on the observation that an educational achievement test showing that everyone in a class has mastered the curriculum objectives by the end of the term is not necessarily a bad test; the lack of variance may simply mean excellent teaching. It may also mean poor measurement. The effect of training seems generally to be an increase in individual differences, not the elimination of them. Absence of variance in scores ought not be confused with absence of variance in the underlying variable being measured by them.

Quite apart from theoretical issues, there are practical reasons for seeking variability in test scores. One is that levels of mastery cannot be identified without individual differences in scores. Another

is that one can never be sure that low variance does not simply indicate an excessively easy test. Of course, it may be useful to minimize variance within groups while maximizing variance between groups. As Millman and Popham (1974) point out, this is a question of validity, and of criterion-related validity at that.

However, if this is one's essential purpose in testing, then the objective of test construction should be a bimodal distribution of scores. In the ideal case, the discrimination values of the items is said to be about .5 (if half of the total sample is classified as masters and the other half as non-masters), and the average item correlation is 1.0. If this were the case, of course, everyone would achieve either a perfect score or a zero score, and a single item would have done as well as the total test. Since the ideal is never achieved, different items represent replications, and the more realistic goal is to obtain as little overlap as possible between the distributions of masters and non-masters.

This seems to be, however, a shortsighted objective in work sample test construction. For most purposes, the distribution of scores should probably be somewhere between a rectangular and a normal distribution. Such a distribution contributes to versatility in test use so that the same investment provides a test for identifying different levels of mastery, for validating aptitude tests, and for diagnosing organizational ills as well; moreover, its use can continue even if performance standards change.

PSYCHOMETRIC CONSIDERATIONS

Work sample tests tend to be reasonably objective. However, their objectivity, as defined in the preceding report of this series, can be

spoiled by distortions in response, characteristics due to restrictive or ambiguous format, poorly motivating testing conditions, or unreliable scoring. Where observers are used, scoring procedures should be defined so completely that high levels of interscorer agreement, or conspect reliability, can be achieved.

It is useful to distinguish the reliability of measurement from the reliability of classification. If the test is to be used for classifying people into those who are certified and those who are not, or into multiple categories, then the reliability of the classifications achieved is an issue of importance. It is, however, an issue independent of the reliability of measurement as such. More will be said on this in the next major section; at the present time, the emphasis is on the reliability of measurement.

That is, the emphasis is on attempts to assure, at the very least, a minimal effect of random errors of measurement. This is the essence of the classical concept of reliability: the extent to which a set of measurements is free from random error variance.

The classical definition of reliability is minimal. Where one wants to generalize the inferences from test scores to "real-world" inferences, assessment of random error of variance yields a necessary but insufficient evaluation. A more important and more general concept is freedom from irrelevant sources of variance, which is the classical concept of validity. There are systematic errors of measurement as well as random errors; in the construction and evaluation of work sample tests, one should be particularly alert to sources of systematic error.

Moreover, one should be alert not only to sources of error

variance, but also to sources of error in individual scores. Since the score on a work sample test is likely to be interpreted in terms of a standard or underlying scale, the test may be useful even in the isolated case where the proficiency of only one person is measured. A satisfactory measuring instrument could be used only one time, for only one person, yielding a score interpretable in terms of test content. This can only happen, however, if the tester can evaluate the degree to which that one use of the test is free from irrelevant kinds of error, even in the absence of a group for which to compute variance.

PREVIEW

In the sections that follow, procedures will be proposed for determining domains, establishing test content, scaling test components, and studying the generalizability of scores. In each section, the different implications for tests of job knowledge, literal work samples, or abstractions of literal work assignments will be discussed. All of this will be offered for the "ideal case" — not for routine testing. In a given testing situation, some of the recommended procedures will be superfluous. In some situations, certain of these procedures will not be feasible.

Despite the differences between generally idealistic prescriptions and the realistic requirements of specific situations, there is a value in being unabashedly idealistic: it provides a conceptual standard against which one can assess the importance of deviations from the ideal in real cases. Without such a conceptual standard, the gradual progress toward improvement — which is an attainable ideal — is unlikely to occur.

DOMAIN DEFINITIONS

Ordinarily, the terms content universe and content domain are used interchangeably. In these reports, however, they have been distinguished; a domain is defined as a not-usually-random sample of a universe. The distinction is made because it offers a useful guide for thinking about judgment processes. It is not intended as a necessary and formal requirement for effective content sampling; nevertheless, the procedures implied by the distinction seem useful, particularly in job analysis.

Although job content domains and test content domains have some different elements in their definitions, both definitions begin with job analysis. The analysis of a job into its component functions can take many forms; the approach outlined here, including the use of the results of the analysis, is suggested primarily as a procedure assuring job relevance of the final test specifications. It begins with a global definition of a job content universe and ends with the specifications for test construction.

Three terms require definition:

1. A component task is a preliminary statement, in rather broad terms, of a major activity, task, or responsibility of the job. It may be an appropriately formalized sentence such as " (Takes action) in (setting) when (action cue) occurs, using (tools, knowledge, or skill) ."
2. A task element is a simple statement describing a detail of the component task; it may describe a movement, a judgment, a source of information, or some other aspect of the broader task. Task element statements may be arranged in the form of a checklist; the same task element may be part of more than one component task.

3. A task category is an empirically-identified group of component tasks sharing a common general pattern of task elements.

The precise nature of the formal sentence may vary according to the nature of the job. For example, Wheaton, et al. (1978), in their study of Tank Gunnery Tests, used a formula somewhat like this: "Given _____ equipment and _____ target under _____ conditions, a tank crew in _____ position will open fire within _____ seconds of the alert element of the command, and neutralize the target within _____ seconds using no more than _____ rounds." Wheaton, et al. described such statements as "job objectives." Their terminology grows out of formal training for tank crew members and follows the language used by educational measurement specialists in writing about criterion-referenced testing. It is commonly asserted in that literature that the unit of analysis in content-referenced measurement is the instructional objective.

The broader term, task component, has been chosen here because content-referenced tests are used for many purposes other than assessing the outcome of instruction. Whether the job incumbent brings the ability to carry out a particular task with him, learns it in formal training, or picks it up on the job is of relatively little consequence in defining the nature of the task. To be sure, in nearly all statements of component tasks are statements of things the incumbent is expected to be able to do; in that sense they are objectives. The present discussion, however, will attempt to evade the overtones of instructional objectives by referring to component tasks.

One component task for an electrical appliance repair person might be "Repairs television set in customer's home at customer's request, using tools in portable kit and knowledge of circuits in that

model and general electronic knowledge". A task element within that component task might include checking vacuum tubes for conformity to specifications. A given task element, such as checking tubes, may show up in other component tasks; it may, for example, be necessary to check tubes in servicing certain electronic air filters. Another task element, "disconnect line cord," would apply to virtually every repair job that he may undertake.

A subjective analog of the task category is the Army's "duty module concept" (Duffy, 1976). A closer example of what is implied by the term task category is provided by Wheaton, et al. (1978) in their cluster analysis of job objectives or, to use the terminology here, preferred task components.

It should be understood that the cluster analysis identifying a task category is not a cluster analysis of checklist items. In the language of Tryon and Bailey (1970), the analysis of checklist statements is a V-analysis, a clustering of variables. The identification of task categories, on the other hand, uses what those authors called an O-analysis, a clustering of objects according to common profiles. The objects, in this case, are component tasks; component tasks with the same profiles of task elements are, for practical purposes in testing, essentially similar tasks, and they may as well be assembled under a common heading.

DEFINITION OF JOB CONTENT UNIVERSE

Job analysis may begin by interviewing job incumbents or their supervisors, alone or in a group, to develop a set of component tasks. The procedure can be expedited by a list of action verbs that can begin the stylized statements of categories. Examples of appropriate

action verbs include "decides," "repairs," "inspects," "assembles," "lubricates," and so forth. The use of key action verbs is more effective if there is a standard set of such verbs, as there is in the Air Force (Foley, 1977). Such a list, however, is not necessary, at least in a group interview, because the crucial list for a given job can be developed on an ad hoc basis. Once the verb for the formula sentence has been selected, the other blanks can be filled in relatively easily, and the first sentence is the hardest.

A job with a large number of component tasks may require a second round of interviews to verify the information obtained in the first. The process of verification is probably less concerned with checking the accuracy of the earlier statements than with stimulating the development of additional ones or of combining those for which the differences are obviously trivial.

When the component tasks have been completely identified, task element statements may be written and assembled into job description checklists. As a general rule, a task-oriented checklist provides a more direct set of specifications for work sample test construction than does a worker-oriented checklist. The latter is very useful in identifying predictor variables, training content, or the essential processes comprising a component task, but it seems less useful in defining the principal activities of the work sample.

Each component task may require several kinds of task element statements. Small panels of job incumbents or supervisors should probably do the initial writing. It may be better if they work as a committee rather than individually if the group activity will stimulate thinking. They should write task element statements under a number of different headings. One heading might be sources of information

or material used in carrying out the task. Another heading might involve work processes: physical activities, perceptual judgments, or cognitive tasks expressed as specific decisions or judgments to be made. Specific actions or accomplishments can be another category, and still another might include prerequisite knowledge utilized in deciding on courses of action or carrying out required actions.

Despite repetition, elements should be listed under each component task. Each statement should be presented with a response scale for the evaluation of the importance of the elements, their duration or frequency of occurrence, percentage of total work time or level of skill they require, or any other scale the panel of experts deems appropriate. The experience of the present writer is that it makes very little difference what scale is used to describe the task elements since the various scales correlate very highly. The important consideration is that members of the panel themselves, and the workers they represent, will feel comfortable in using the scale chosen.

After a pilot study to assure clarity and completeness, the job description checklist should be administered to a large sample of incumbents or supervisors. The sample completing the checklist should be representative of potential diversities within the organization. If the job appears, for example, in regionally scattered installations or in installations varying markedly in size, then incumbents from each of the regions or each installation size should be surveyed. The survey can usually be done by mail, although interviews may be helpful if the checklist is complicated.

Data from the survey should be analyzed to identify the most common elements making up individual component tasks across sample characteristics such as region or installation size. Some task

elements may be an actual part of the component task for only some respondents. Elements of a component task that are idiosyncratic to individuals or subgroups do not really define the task.

One might do a cluster analysis of all task elements and determine cluster scores for each component task by averaging across respondents. Each component task will then have a profile consisting of the same number of points as clusters in the checklist. If there has been sufficient uniformity of checklist items from one component task to another, these profiles can be compared in developing task categories. Clusters of component tasks with similar profiles identify task categories. The job content universe consists of the set of task categories, each defined by its own characteristic profile of task elements.

In developing the Tank Gunnery Test, Wheaton, et al. (1978) identified a list of 266 job objectives (component tasks) and 114 behavioral elements (task elements) of those "objectives." Assigning a dichotomous classification of either one or zero for each cell of the matrix (based on a restrictive assumption of a first-round hit), indicating whether the behavioral element was in fact involved in the job objective (component task), they cluster analyzed that matrix. The results of the analysis identified 16 relatively homogeneous clusters in terms of the behavioral elements that adequately defined the domain of interest if not the total job universe.

DEFINITION OF JOB CONTENT DOMAIN

The complete list of component tasks, each of which is defined in terms of its own list of task elements, offers a thorough definition of a job content universe. The definition of the job content

universe is redundant and can be simplified without serious information loss to whatever degree component tasks may be clustered into task categories. Further redundancy exists because certain task elements are repeated far more frequently than others across the various categories. A random sample from the job content universe would, therefore, produce an unnecessarily redundant examination. Moreover, some parts of that universe may be trivial for the purposes of testing.

The definition of a job content domain is a matter of sampling from the job content universe, but it is by no means a matter of random or representative sampling. The sampling strategy should fit predetermined objectives. In the gunnery tests, for example, Wheaton, et al. developed a sampling strategy based on an "index of generalizability." Generalizability in this case is a function of the number of behavioral elements (task elements) in one component task that are also included in one or more other component tasks. They had decided that there would be proportional representation from each of the task categories according to the number of component tasks they contained. If proportionality suggested that one single component task would be chosen from a given category, it would be the one with the largest index of generalizability. If proportionality required two component tasks, then those with the largest indices of generalizability were chosen, and so on.

Several judgments such as these are made in defining a job content domain, and they should be made by a panel of experts consisting of job incumbents, their supervisors, or both. (Systems experts, industrial engineers, or others may be useful for some panels.) For a test designed to certify competence, such a panel needs to determine which kinds of task elements are the essential or critical elements in the various task categories. Decisions made at this point will

greatly influence subsequent decisions whether to develop a test measuring hands-on performance, a written test of job-related decisions and judgments, or a job knowledge test. These decisions will not necessarily come easily; the test developer may expect sometimes acrimonious debate between those who believe that nothing matters but the real results of work and those who believe that good results are attributable only to dumb luck in the absence of well-informed judgment. Somehow, however, the different opinions must be reconciled and a consensus reached about the criticality of the task elements. What seems to be important in making these judgments is that strategies for selecting the sample of component tasks or task elements from the total universe be developed clearly, be accepted by the panel of experts who must make subsequent judgments, and be reasonable in the light of the objectives of work sample testing.

DEFINITION OF TEST CONTENT UNIVERSE

In his chapter on test validation, Cronbach (1971) referred to the "universe of admissible operations." Later, in the monograph on generalizability theory, he and his colleagues referred to the "universe of admissible observations" as the basis for domain definition (Cronbach, et al., 1972). The terms are doubtless interchangeable, but a possible difference of emphasis that makes a bit of word play instructive for the definition of work sample universes and domains.

Operation seems to imply doing something; admissible operations might be the acceptable things test builders do in providing the stimulus material or that test takers do as responses to that material -- the stimulus-response content of the test. Observation seems in addition to imply noting and evaluating what the test taker does -- the scoring content of the test. For work samples, at least, it seems the more inclusive term.

The Universe of Admissible Observations. It is very difficult to imagine a work sample performance test that does not involve measurement by direct observation. The measurement of such a test is the result of an evaluation either of the procedures used in carrying out the requisite tasks or of the evaluation of the results of task performance. In either case measurement depends on some form of observation or an observational aid. It is unlikely that credence could be given to descriptive accounts by job incumbents about the work processes, not because of possible faking so much as because jobs tend to become automatic. People who know how to do a job well simply do not know for sure what they do. If it is a product of the work that is to be evaluated, some form of inspection is needed. Inspecting is careful observing. It may be aided by various kinds of physical instrumentation ranging from nothing more complicated than a ruler to massive equipment for testing stresses or breaking points, but the final responsibility for judgment in the inspection of a product rests with the observer; ultimately, therefore, the measurement of job proficiency through the evaluation of the products of work is a form of direct observation.

A definition of a test content universe must specify both the stimulus materials (the assignments or questions) and the general form of probable or permissible responses (performance or answers). If it is a work sample test, the stimulus-response content must be drawn from the job content domain. The scoring content, however, does not exist in the job content domain, at least not ordinarily. It must be an added factor, and the emphasis on watching and scoring implied by speaking of observations makes it useful to define a test content universe as a universe of admissible observations.

In their generalizability monograph, Cronbach et al. (1972) point

out that any observation can be described in the context of certain conditions. These may include such considerations as the nature of the task, the environmental setting in which it is performed, the time of day, or the level of external control over performance among many other possible considerations. Each of these is a facet of the observation, and there can be limits to the range of conditions within each facet of interest. In Army testing for skill qualifications, for example, environmental hostility can range from nearly none at all to combat conditions. One may decide on either technical or social bases that the admissible or acceptable set of environmental conditions for purposes of testing people may be limited to simulations of certain features of the combat condition.

The task components and some other facets of the universe of admissible observations are specified in the job content domain. These are not enough for testing. Other facets or conditions must be added to produce a generalizable set of scores. This is why a test content universe or domain cannot be equated with a job content domain. The added components must be added with careful, informed, and systematic judgment by qualified experts, and attempts should be made to consider, even if ultimately to reject, any potential element in the universe of admissible observations.

Item Forms. A useful guide to the development of a universe definition is the concept of the item form (Hively, Patterson, & Page, 1968). The research reported by Hively et al. described item forms for a test of basic arithmetic skills — clearly a finite universe considerably smaller than the universe for most work sample testing. However, the stylized sentence suggested for the identification of component tasks can be used to develop item forms for work samples. Additional elements are needed, including elements for defining the

circumstances of observation, the methods of scoring, and some given conditions to be assumed. One item form for the electronics repair work might, for example, follow the form: "Given (diagnostic data) about a malfunction in (product), and given (conditions), candidate must (locate and replace) malfunctioning (part) with the work or response evaluated by (method of observation or scoring)." That particular item form may suggest many specific items. With all blanks filled, a change in any one of them defines the content of a new item. If the malfunctioning part is a condenser, replacing it may require soldering, and the solder connections can be evaluated either by an inspector's rating or by measuring the current flowing through the connection. Changes in diagnostic data or conditions can result in quite substantially different items, differing not only in content but in difficulty. Different item forms may require the candidate to develop diagnostic data, or to make judgments necessary for subsequent steps in a process, or simply to answer questions.

The number of possible item forms is obviously very large, even for a very simple job; more complex jobs might require, if not an infinite number of item forms, a prohibitively large number. However, a test content universe can be said to have been defined when the rules for generating item forms have been specified, even if no actual examples exist. The rules themselves, if fully stated, can identify the nature of responses that can be obtained and the variety of ways those responses can be evaluated.

DEFINITION OF TEST CONTENT DOMAIN

By establishing some item forms and rules for the generation of others, a test developer and his panel of experts will have shown the plausible limits within which testing can conceivably be done. It does

not follow that all plausible item forms are worth developing. The panel of experts will probably dismiss some item forms out of hand as too costly or otherwise impractical. A process of elimination may be necessary to select from the universe of all possible observations a restricted subset for actual test development.

In one job analyzed by the writer, it was found that one task category consisted of reading material of varying complexity to obtain information fundamental to carrying out other duties of the job. The universe of admissible observations could be restricted to conventional reading comprehension testing. Facets included, in addition to the stimulus content material, facets of item format (multiple-choice, true-false, arrangement items, essay items, fill-in items, etc.) and facets for scoring responses (differential vs. unit weights, use of machine scoring vs. independent observers rating open-end responses, etc.). The operating decision was made, largely on the basis of the practical considerations of the number of people to be tested and the time period within which the testing had to be completed, to measure comprehension with conventional multiple-choice items following sample passages to be read. Rules for sampling material for the stimulus passages were established, and multiple-choice items -- a very conventional "item form" -- were chosen for development. The test content domain, therefore, consisted of passages to be read, sampled according to the rules established, and multiple-choice questions covering that material. The deficiency in the definition of the test content domain in that particular instance was that the rules for determining numbers of items, difficulties of items, and other related matters were never specified.

Test Specifications. In any case, with greater or lesser precision, with greater or lesser representativeness of the possible

universe, with greater or lesser ambiguity, the panel of experts, under the guidance of testing specialists, will arrive ultimately at a set of specifications for test development. These will be in part content specifications, in part format specifications, in part response specifications, in part scoring specifications. There may also be some structural specifications for the inclusion of items; although the work sample test is intended to be a content-referenced test, the panel may specify limits of conventional item difficulty levels or discrimination indices.

TEST DEVELOPMENT

If the specifications call for a conventional job knowledge test, test items must be written. If the specified item forms require that something more than rote memory is to be invoked in testing for job knowledge, the items must be written to be challenging and to require thought while taking the examination. For hands-on performance tests, the elements (they, too, are items) of the tasks should also tap fundamental rather than superficial. The development of really good items is the foundation for either kind of work sample testing.

QUALIFICATIONS OF ITEM DEVELOPERS

Item writing or development is an art, and it is also hard and highly specialized work. Large test publishers or major civil service jurisdictions maintain on their staffs full-time, professional item writers; paradoxically, in highly specialized fields, the people most likely to have the necessary knowledge and full understanding of the implications of that knowledge are the people who have worked in that job, not those who have worked as professional item writers.

Wesman (1971) identified six points describing the combinations of abilities necessary to write good test items; they apply also to developing good exercises to fit performance test item forms:

1. The item developer must have full mastery of the subject matter of the test; full mastery does not imply merely acquaintance with basic facts and principles but the understanding of them and their implications. This implies awareness of popular fallacies and misconceptions.
2. The item developer must have a clear understanding of the objectives of the test and of the reasons for testing. In educational testing, this implies an understanding of curricular and educational objectives as well as the specific test objectives. In developing a work sample test, it implies understanding of organizational values and of why doing certain things well on the job may be more important than other things that are relatively trivial but easier to put in a test.
3. The item developer must understand the characteristics of the people for whom the test is constructed. This means not only an awareness of the examinee's anxieties but also implies an awareness of the ignorance or clumsiness of potential examinees that can lead to mistakes, including acceptance of plausible wrong answers.
4. The item developer must be excellent in the use of language. This requires not only a useful vocabulary but skill in arranging words so that their precise meaning is inescapable. It applies as much to instructions as to verbal test items.
5. The item developer needs to understand specific techniques of item writing, including familiarity with different types of test items, their possibilities, and their limitations. Wesman points out that skill in item writing means more, however, than merely an understanding of item types. It requires imagination and ingenuity to create the kinds of situations, sometimes on paper, that will evoke expressions of knowledge. It requires similar imagination and ingenuity to abstract critical exercises from long and complex tasks.
6. Item developers who are skillful in one context or test type may find themselves less skillful in others; they need to learn their own skills and to collaborate well with others whose special strengths are complementary.

It seems clear that the combination of the professional item developer and the job incumbent is essential to the development of a job knowledge test. This combination is not likely to be found in any one person; test specialists must therefore train potential item writers in the skills of item development. A panel of experts is needed who have performed the job and know it well, either as incumbents or supervisors.

The first orientation meeting for panel members should clarify at the outset the purposes of the test to be developed and the general principles and values that guide the test construction enterprise. This should be followed immediately by a general instructional in item forms, item types, and item development. It is unlikely that imagination and ingenuity can be created in a brief training period, but they can certainly be stifled in such training if signs of them are not rewarded. A standard textbook, or, indeed, Wesman's chapter on item writing, can provide text material for such training; others (e.g., Boyd & Shimberg, 1971; Jones & Whittaker, 1975) can serve for examples of hands-on test items. As each item type is discussed, panel members should be encouraged to try to prepare the various kinds of items to fit the existing specifications, and examples of ingenious items should be brought to the attention of all.

Such training may seem largely superfluous if the test specifications have identified useful item forms and have specified in detail the test content. Even with such detailed specifications, however, item developers who understand their options well can be expected to be more creative in the invention of items within the item forms.

In short, item developers must be qualified, first, in terms of job experience and second, in terms of special training in the

techniques of developing test items. A third qualification to be effective in developing items is motivation. The item developer must accept the purposes of measurement, believe in it, and have a desire to contribute to the development of an effective test.

ITEM POOLS

Although test specifications specify the number and kinds of items to be prepared and the content balance among them, the item developers should produce a surplus of potential items. Wherever possible, they should develop a pool of items at least 2 1/2 times the number needed for each item form. Such an item pool makes it possible to develop two randomly parallel forms of the test (Cronbach, 1971) without having to resort to items of questionable quality. Having two forms has a substantial number of benefits, not the least of which is the benefit in retesting people without compromising the security of the first form. However, the main reason for recommending two forms is that the evaluative data analysis to be recommended often requires them. The correlation of scores on these two forms, if nothing else, is an indication of the relevance of item content to the content domain and of the degree to which the content specifications of the test were clear enough to produce essentially similar instruments.

Job Knowledge Items. Job knowledge tests are not necessarily written tests. Comer (1971) described a trade test for the automobile service trade which used actual specimen parts from serviced automobiles as the source of items. The examinee could look at, handle, feel, or even smell each part. For each one, he was asked to (a) identify it, (b) describe its condition, (c) decide why it failed, and (d) decide whether it could be reused. The rationale is that an

uninformed person would not know what to look for in examining a part. "He would not recognize pertinent symptoms, nor distinguish between normal, harmless scratches and discoloration, and true damage" (Comer, 1971, p. 50). For each part, the examinee filled in a blank with the name of the part and checked one alternative each describing condition, reason, and prognosis for further use.

The "items" in this test might be considered either the automotive parts examined or the questions asked about them. Since there were ten parts and four questions about each one, the test has either ten items or 40. For developing the item pool, it is useful to consider each question an item; for item analysis, there may be only ten items. If there are contingencies among the four questions in each part (e.g., if knowing the nature of the damage depended on proper identification of the part), then neither conventional methods of item analysis nor latent trait analysis can properly be used.

There are no similar uncertainties in most job knowledge tests; they typically are measured through the use of multiple-choice or other more or less objectively scorable, clearly independent test items. Nothing is needed here about suggestions for writing such items since there are many useful sources available (e.g., Ebel, 1972; Wesman, 1971). The item developer should be warned, however, that, although item writing looks easy, the only easy thing about it is an easy superficiality.

The problem is reflected clearly in the study of occupational licensing practices by Shimberg et al. (1973). They attributed the many flaws in written licensing tests to four categories:

1. Few of the licensing boards did any planning for the test

beyond a vague outline, and some did not even have an outline.

2. Many state and local boards have a compulsive preference for essay tests; essay items are often ambiguous and scoring procedures are usually unreliable.
3. Where multiple-choice items are used, many licensing boards prepare items which are answered correctly most readily by people who have memorized the review books from which they were drawn.
4. There is little evidence of any concern for statistical item analysis or any other item evaluation procedures; there seems to be an attitude that once the test items are written, the problems of test construction are over.

Items in the Total Job Test. Jones and Whittaker (1975) refer to "total job" tests. In these, the examinee is simply doing the job (or a thorough replica or simulation of it); the only thing that identifies it as test performance is that the work is done under standardized conditions, that the performance is systematically observed, and that the observed performance is evaluated directly. What are the "items" to be developed for such a test?

Perhaps the ultimate "total job" test is a carefully designed probationary period in which the worker is systematically rotated among different assignments or work stations. Each stop in the rotation may be an item or, alternatively, a subtest in which the items might be units of time, such as production or scrap per hour, or points on an observer's or inspector's checklist.

This is a limited form of testing. Only relatively routine, short-cycle jobs fit easily into this frame. For most jobs, samples of the work to be done are practical necessities. Such samples can for convenience be divided into those in which the sampling is rather direct and obvious and those in which the sampling evolves from a

process of abstraction resulting in a test that is quite different in appearance from the actual job.

Items in Direct Work Samples. The item forms established from test specifications determine the nature of the items. Each item form identifies a task to be performed and the circumstances in which it is to be performed; that task is an item. As in Comer's (1971) test for the automobile service trade, however, there is always some uncertainty about whether it is the task as a whole or the component scorable parts of it that are the literal analogs of test items. The solution to the uncertainty, as before, seems to be in determining whether there are contingencies in the finer divisions of scoring performance that will interfere with effective item analysis.

Foley (1977) speaks of "scorable products"; Maier, Young, and Hirshfield (1976) refer to "scorable units." In a literal sense, any element in performance, product, or observation that can be graded, rated, classified, or otherwise scored is a de facto item. To identify scorable units, the panel of experts may need to develop successively smaller divisions of tests, subtests, item frameworks, and independently scorable units. It must determine precisely what events or activities or attributes of products can be observed and how they can be evaluated.

A few examples will identify the variety of work sample items that can be invented:

1. In marksmanship performance tests, firing a specified number of rounds from a specified position a specified distance from the target is an item (Wright & Mead, 1978).
2. In a dental hygienist's work sample, performance is observed

by an observer with a checklist of specific steps in each of several task sequences; each step is an item rated by the observer (Boyd & Shimberg, 1971).

3. In an offset printing work sample, items are statements in lists of violations of good offset press operation and violations of safety rules; each violation noted is checked (Boyd & Shimberg, 1971).
4. In a test for truck mechanics, one part of the battery is the discrete task, "remove and replace clutch plate." Nineteen penalty scales for specific examples of poor performance, under five headings, were rated for degree by an observer (Jones & Whittaker, 1975).
5. A metal lathe operator may be asked to make a taper plug (not because it is an actual piece, but because it requires so many kinds of turning) according to engineering specifications. Items include (usually dichotomously) the match of the product to eleven specified dimensions and could include ratings of both smooth and knurled surfaces (Jones & Whittaker, 1975).
6. In a tank crew gunnery test "engagements" are items; the 28 engagements varied according to gun used, time of day, responsible crew member, mode of firing, target type, target range, and other facets of the exercise. Items were scored as hits or misses, either on first trial or on second trial (Wheaton et al., 1977).
7. The REALTRAIN battle simulation has no clearly identified a priori items; however, a posteriori items can be derived from a net control station data recording sheet identifying casualties claimed and confirmed (Shriver et al., 1975). It is worth noting that the exercise in its entirety, as are many other simulations, actually constitutes a single item.
8. In a performance test for radar electronic maintenance technicians, 81 problems involving use of test equipment, adjustment or alignment of test equipment, and many others led to a total of 133 "scorable products" to be completed satisfactorily (Foley, 1977).

Most work sample tests are reported with no clear identification of items or scorable units of measurement. Apparently, test developers

have given little attention to the concept of an item of measurement. Many large-scale, single-item simulation exercises, carefully and elaborately planned, are excellent management or training tools -- like REALTRAIN, which was designed explicitly for training purposes -- without being particularly good measures of anything. In nearly any measurement by testing, items of measurement are important in one of two major roles: either as indicators of units on a fundamental scale of measurement or as replications, as in the several items of an achievement test. In many work samples, the scored items are contingent items; "passing" the item depends upon prior success with a previous item. Such items are certainly not replicates; neither will they form a scale without very careful planning. Treating overall performance on an exercise as a whole, or as a single item, risks seriously unreliable measurement.

It is here asserted that better measurement will be obtained from direct work samples only when item developers concentrate on identifying and developing work sample analogs of independent test items.

Items in Abstracted Work Samples. The distinction between a direct and an abstracted work sample is no more than the distinction between ranges on a common continuum; they are not distinctly different kinds of tests. The difference is that direct work samples have at least the appearance of "real" work assignments. They do, of course, involve some abstractions; the machinist who is required to make a taper plug, for example, may never again have a specific task assignment which requires all of the specific kinds of cutting required by the test -- but assignments from a variety of "real" jobs are abstracted from them and put together in one "artificial" task assignment.

Abstracted work samples may seem less like the real job than like a test, but they nevertheless clearly sample the activities and the skills of real jobs. An example is the typical In-basket Test. An In-basket usually involves inventing a make-believe organization with imagined organizational relationships, problems, and required decisions. Within this imaginary framework, pieces of paper are created to be similar to papers found in actual in-baskets of real people in real organizations. The examinee must pretend that he is really in the make-believe world of the imaginary organization, that he is new to the organization, and that he is working at night with no one else around to give him needed information not in his packet of test materials. Yet the persistent reaction of people who take such tests is that they are "realistic" — that is, the In-basket Test, if reasonably well constructed, poses such real-to-life problems that its artificial aspects fade into insignificance as the examinee becomes more engrossed in its basic reality.

Abstractions may maintain the "hands on" character of the direct work sample while seeking independently observable performances as items. To illustrate the problem, assume the home workshop project of making a lamp out of a bottle set on a specially turned wooden base. Four independently acquired skills are used: turning the base, cutting the bottom off the bottle, assembling the various parts, and attaching the wiring. A direct work sample would give the examinee the necessary plans, materials, and tools and set him to work. The results might be scored in terms of whether the lamp works, whether it is steady on the table, or ratings of various aspects of its appearance. An abstracted work sample, in contrast, might be four quite different tests. One might be a turning test, in which several wood turning skills could be demonstrated without designating any product as a lamp base, the second might be a similar test for bottle

cutting, the third an assembly test in which a previously turned base and bottomless bottle were provided along with the other lamp parts, and the fourth the simple wiring of an assembled lamp. At no point is one step of the lamp-building procedure dependent on how well the previous steps had turned out, nor is it necessary that the worker have a lamp in mind as the final product while taking the first two tests. All that is necessary is that the component tests (or subtests) tap the necessary skills.

Abstractions may also minimize or even abandon the "hands on" feature. Job knowledge tests are usually abstract work samples. That is, the knowledge components of the job content domain have been abstracted to form the nucleus of the test content universe and domain. The result is in no sense a full work sample, but it can be considered a partial work sample — a sample of what expert judges considered an essential component of the total job. Comer's test for the automobile service trade is an example. A direct work sample could have been developed that would require the demonstration of skill in removing the parts central to the test, diagnosing them, and taking appropriate corrective action. The essence of corrective action, according to Comer (1971) is the knowledge one brings to and derives from examining the part and deciding why it might look as it does.

What is here called an abstract work sample may be very much like the test Foley (1977) termed a symbolic substitute for a performance measure. What is intended by the term is, however, more than symbolism. What is intended is literally pulling from the whole an essential element — an abstraction — on which the abstracted performance can be used to infer performance on the whole.

The process of inference clearly identifies the abstraction as a

test; it places the focus on the score and on the necessity for an empirical validation of the inference. It also places focus on item definition as the process of test construction; the item form is seriously taken as a prescription for an item, e.g., "Given all necessary parts and ordinary hand tools, candidate assembles lamp ready for wiring."

Requirements for Scorable Items. There are many kinds of work sample tests. A multiple-choice job knowledge test is an abstract sample of the job (Campbell et al., 1973). So is an abstraction of decision processes (NCARB, 1976) or of manipulative activities (Comer, 1971) — or, for that matter, a standardized observation of a full cycle of actual job performance for a specified time period. Whatever the nature of the sample, it can usually be analyzed into component parts, and performance on these components can usually be scored, graded, or rated; each component is therefore a test item which, aggregated according to a specific rule, yields an overall score or subscore. These items may be as complex as a battle simulation or as simple as a brief time span.

What is required for such items during the development of an item pool is that each be independently scorable and that each be traceable to its origin in the job content domain.

ITEM ANALYSIS

Once a pool of items has been prepared, item analysis is necessary both to identify poor items and to verify the match of actual items written to the defined test content domain. Verification requires the systematic collection of judgments; other kinds of poor items can be identified by conventional item analysis.

Retranslations. The judgments concerning the fit of items to the test content domain can be obtained in a variety of ways. One possible method is an adaptation of the "retranslation of expectations" proposed by Smith and Kendall (1963) for the development of behaviorally anchored rating scales.

In brief, the first step in the Smith-Kendall procedures requires a group of experts to identify the dimensions along which ratings of performance should be made. An analogy to content sampling is the identification of components of the job content domain. A second step is to write behavioral statements for different levels of each dimension, for which the parallel step in test construction is writing test items. A third step convenes an independent group of judges to allocate each behavioral statement to the dimension from the original list which it fits best. The "retranslation" analogy is the practice of translating a passage in English into a foreign language, and then retranslating the foreign language passage back into English. If the two English versions — the original and the retranslated — match, it is assumed that the foreign language translation was satisfactory. By the same analogy, if items are developed to fit one or more components of a defined test content domain, then an independent panel of judges should be able to allocate each item to its proper components. If there is no consensus about the fit of an item in the content domain, it is identified as a poor item; where there is consensus, the appropriateness or relevance of the item for its intended purpose in the domain is verified or established. If there are components of the overall content domain for which the items written have not "retranslated", then deficiencies in the item pool are identified.

Content Validity Ratio. Lawshe (1975) presented an equation for computing a statistic describing the degree to which an independent

panel of knowledgeable experts considers the knowledge or skill measured by an item to be essential to the performance of the job. He called it the content validity ratio. For a variety of reasons, including the fact that the technique does not identify deficiencies in content sampling, this writer prefers to call it a job relatedness ratio; it provides an index number for evaluating the relevance of an individual item for job performance.

The job relatedness ratio is computed by the formula:

$$jrr = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

in which n_e is the number of panelists classifying the content of the item as essential to satisfactory job performance (as opposed to "useful but not essential" or "not necessary"), and N is the total number of panelists. The ratio is a direct linear transformation from the proportion n_e/N , but it has certain advantages other than simple proportion. If less than half of the judges consider an item essential, the job relatedness ratio is negative. If exactly half say that the knowledge measured is essential, the ratio is 0. If all say that the knowledge is essential, the ratio is 1.00. In short, the range of the job relatedness ratio corresponds to that of a correlation coefficient and may therefore be interpreted in fairly familiar ways. Another advantage is that the minimum values for the job relatedness ratio to be statistically significant at different panel sizes have been reported by Lawshe.

The ratio does not identify components of a test content domain omitted from the item pool judged. This seems to be a relatively

trivial matter. If the job analysis defining the job content universe and domain has been thoroughly carried out, and if the judgments extending the job content domain to the test content domain and to test specifications have been thorough, then item development to meet those specifications should provide a rather complete coverage of the job domain. As a matter of fact, any component of the domain that systematically loses items on the basis of low job relatedness ratios probably should not have been in that domain in the first place. In such a case, the original panel of experts should certainly give serious consideration to redefining the domain without that component or to writing new items that might have a better chance of being judged essential.

Thus the job relatedness ratio offers real advantages in item analysis. It provides a procedure for systematizing and documenting the judgments of a panel, it provides an index number that can be readily interpreted, and it provides a record of the judged importance of individual items that can serve as evidence of the job relevance of individual items in the case of litigation.

Index of Item-Objective Congruence. A combination of procedures like those described above characterizes an overall statistic reported by Rovinelli and Hambleton (1977). To place their index in the context of personnel testing, the reference to curricular objectives may be taken either as a reference to component tasks in defining a job content domain or universe or to the task category defining the job content domain.

Each judge in a panel evaluates each item in the pool for each of the components. Evaluation is expressed on a 3-point scale of +1, 0, or -1, indicating either a clear allocation of the item to the component, indecision, or a clear belief that the item does not fit that component.

The computation of the index is based on a data matrix in which the judgments of the individual content specialists may be considered the columns and the objectives or job components may be considered the rows. The index of item-objective congruence is then given by the equation:

$$I = \frac{(N-1)\Sigma X - \Sigma \Sigma X + \Sigma X}{2(N-1)n}$$

where

I is the index of item-objective congruence for a specific item on a specific objective for a component,

N is the number of objectives for components,

n is the number of judges,

ΣX is the sum of the ratings for that item by the judges, and

$\Sigma \Sigma X$ is the sum of the summed ratings across objectives.

The resulting index, like the job relatedness ratio described above, is an index number that conveniently ranges between -1 and +1 with the 0 point indicating that all judges are undecided about the allocation of the item. The major feature of this index is that it is not simply a global judgment of relevance, but is a judgment of relevance to specific components of the content domain. Rovinelli and Hambleton recommend that the index be computed for every item for every objective, a recommendation which has the merit of identifying ambiguously assigned or redundant items. Like Lawshe, they make no particular recommendation concerning a minimum value of the index deemed acceptable for item inclusion; this, they say, will depend on experience both with a panel of content specialists and with the use of the index. A disadvantage of this index in relation to the job relatedness index reported by Lawshe is the absence of a significance

test; this is probably not a serious defect; one could easily be developed.

Conventional Item Analysis. The statistical item analysis techniques conventionally used in norm-referenced test construction may also be applied to items in work samples. Whereas the foregoing analyses have been based on a priori judgments of the items, conventional item statistics are based on actual responses to the test items from a specific sample of examinees. The sample should be as representative as possible of the population with which the test is to be used. If the test is to be used for certifying knowledge or proficiency on a job (or denying such certification), then the sample should consist of candidates for certification. This conventional requirement for a representative sample poses some special problems for the item analysis of work sample tests. People applying for certification are probably highly self-selected so that most of the candidates will pass most of the items; that is, there may be little variance either in item responses or in total test scores. Conventional item analysis techniques require at least some variance in both.

The most useable item statistic is its difficulty or easiness level. It is conventionally computed simply as the proportion of people in the sample passing the item (e.g., answering a question correctly). It used to be said that the average difficulty level should be somewhere around .5, a prescription with little value for the developer of a content-referenced work sample test. One might determine the rank order of items in the item pool in terms of their difficulty; the most difficult items and the easiest items can be reexamined for the appropriateness of their inclusion. An item for a critical content component might, if excessively easy, fail to measure adequately the salient knowledge or skill. On the other

hand, an item that is too difficult might also fail to measure it because of ambiguity, fuzzy instructions, or some other defect. The computation of the item statistic, therefore, should not be considered the final evaluation of the item. The panel of experts can be reconvened to reconsider items in the light of their descriptive statistics, and any item which, in its collective judgment, is either too easy or too difficult, may be revised.

The other conventional item statistic is the item discrimination index. If an item correlates well with the total score, it is assumed to differentiate well between those who are knowledgeable or skillful and those who are not. A low correlation identifies items assumed to be poor in distinguishing the certifiable from those who are not.

The absolute value of these correlations is usually trivial since it depends so much on the variance in the available sample. However, the item-total correlations can be used for rank ordering the items.

High item-total correlations are usually considered unnecessary in content-referenced tests; such tests need not have high levels of homogeneity (Cronbach, 1971). However, such tests should have at least some functional unity. Functional unity means that items within a test should somehow hang together; everything in the test should be at least a little bit correlated with everything else. In other words, the item-total correlation need not be high, but it should be positive. Zero or negative correlations identify either poor items that should be revised or replaced or subsets of items that should be independently scored. It is poor measurement when a test score represents an unknowable combination of independent or even negatively related attributes.

SCORING AND CALIBRATION

The conventional method of scoring a conventional test is simply to count the number of items with correct or passing responses. A multiple-choice test may be scored according to a formula, such as the number of items correctly answered minus a fraction of the number of items answered incorrectly (the fraction depending on the number of optional responses available). Each item is a unit. One item, easy or hard, counts just as much as any other item; differential weighting of items is traditionally considered useless with large numbers of items. Hambleton et al. (1978) offered five methods for estimating examinee domain scores or "true proportion correct scores."

Item analysis may, particularly with small numbers of items, provide a basis for differential weights, but it is rarely if ever done. Differential weighting of items in work sample tests is rather common, but it is usually based on a priori judgments of relative importance rather than on item statistics.

Strictly speaking, it is usually improper to refer to a conventional score on a psychological test as a measure of an attribute; it is better to describe it as a sign or reflection of the attribute. Whether the score really can be used as an indicator of the aptitude is a question to be answered as its construct validity.

It has been pointed out that concepts of validity are unnecessary in fundamental measurement. If work sample tests are to be more than reflections of level of performance, if they are in fact to be formal measures of performance variables, then scoring should be more than merely a count of right answers. The score must form a scale with fundamental mathematical properties, at least the property of ordinal

transitivity. The translation of scores into such scales involves item or score calibration. One simple method of calibration will be proposed, based on the logic of what Ebel (1962) called the content standard score. Another one will be briefly mentioned (Anderson, 1976). The major discussion will, however, be devoted to latent trait theory, a well-developed approach to calibration through so-called "sample-free" item analysis.

LATENT TRAIT ANALYSIS

Latent trait theory overcomes the problem of over-reliance on a specific sample. Wright (1968) has demonstrated that calibrating items using the one-parameter Rasch model will yield ability estimates in new groups of people that are independent of the ability distribution in the particular sample of people on whom the test was calibrated. Moreover, estimates of item difficulty level can be made more or less independently of the distributions of difficulty in the set of items used in making the estimates. That is, the item and ability parameters are essentially invariant across samples of people and of items; this, according to Wright, is the ultimate in objectivity in measurement.

The Rasch model estimates only one parameter of the item characteristic curves. This is generally justified on the grounds that three-parameter models are less successful in finding invariance for parameters other than the difficulty parameter, although Slinde and Linn (1978), finding insufficient evidence of invariance using the Rasch model thought additional parameters would help. Item discrimination values are less reliably estimated, and the estimation of so-called "guessing parameters" is often seriously unreliable. Rudner (1977) has demonstrated similar objectivity using a three-

parameter model, but the estimates of the a and c parameters were considerably less reliable than those of the b (difficulty) parameter.

Proponents of different latent trait models seem adamant in their preferences; these reports will not take sides in disputes on the general relative merits of the different models. However, for the purpose of developing a job knowledge test, the application to be discussed first, the three-parameter model will be used as the prototype. There are two reasons. First, it seems to be the more general model, giving more information than is possible with the Rasch model. Second, it provides another form of analysis by which one kind of "poor item" may be identified.

Consider Figure 1. It proposes to show three-parameter item characteristic curves for four items in a given job knowledge test. The items are arranged in increasing order of difficulty so that they may be said to form a genuine scale, at least p_b probability of a correct response. The a parameter, the discrimination parameter, is essentially the same for items 1, 2, and 3. It is markedly different for item 4. For people with high levels of ability, such as θ_2 , items 1, 2, 3, and 4 form a progression of difficulty with item 1 the easiest (highest probability of a correct response) and item 4 the most difficult. For people with levels of ability at θ_1 , however, item 4 is the easiest, that is, it has the highest probability of a correct response.

Alternatively, the problem can be seen by examining the ability level required for a specific probability of a correct response. At p_b , the ability level requirement increases from item 1 through item 4. At p_a , however, the ordinal positions of the items in terms of required ability level is 1, 4, 2, 3. From the point of view of

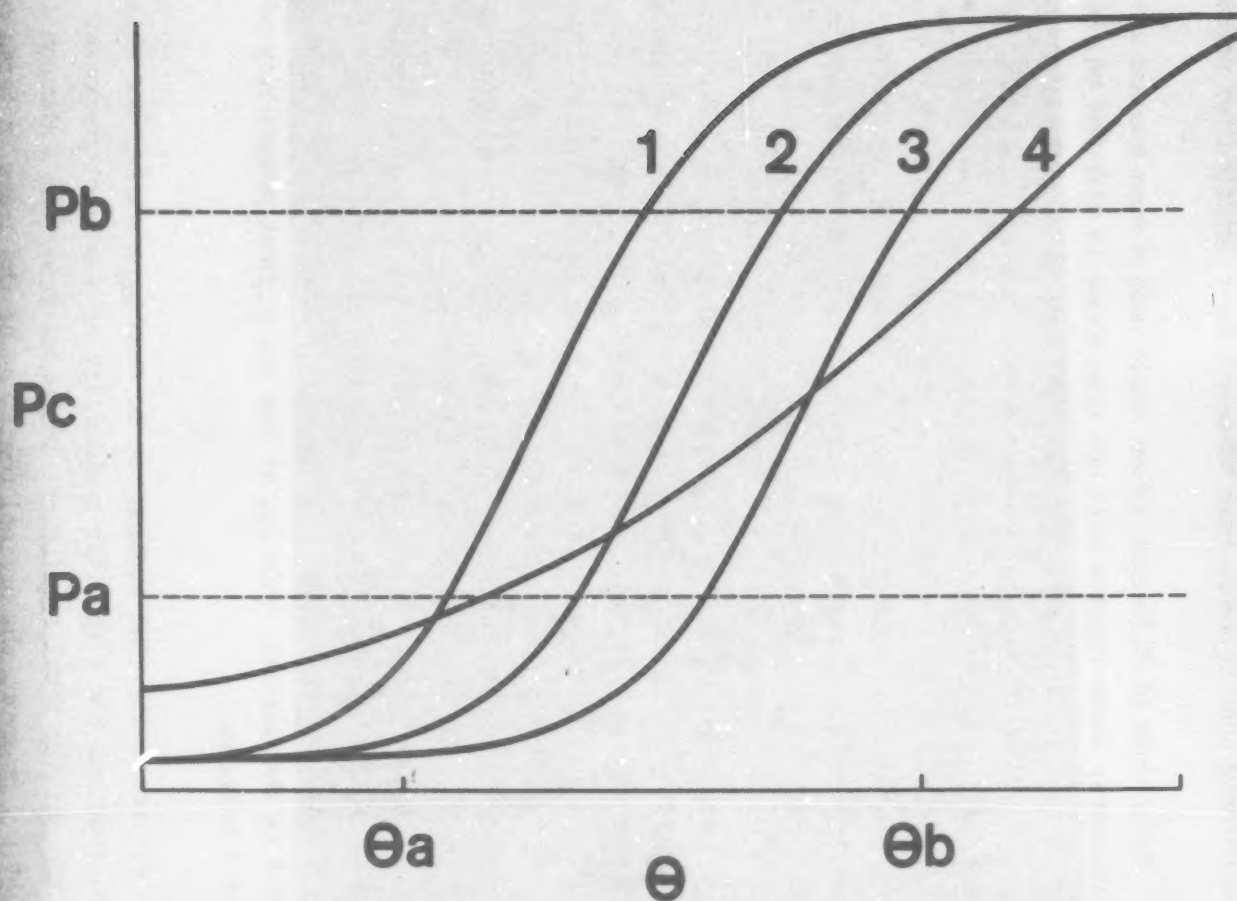


Figure 1. Three-parameter item characteristic curves, for four hypothetical items, showing changing order of difficulty at different ability levels.

formal measurement, the first three items are excellent; they form a true, transitive scale of measurement regardless of who is tested. Item 4 spoils the scale; with it included, the transitivity of item order varies according to ability level.

Clearly, item analysis by latent trait item characteristic curves can not only lead to item calibration but also to identifying poor items from the standpoint of the formal properties of measurement.

Suppose that isolated examples of such intransitivity were found. These would identify items that should either be deleted or revised. Suppose, however, that a clustering of items with similar discrimination values were found and that they were associated with different components of the content domain. Such a finding would be a reason for establishing independent scales for measuring independent components, even though the scales might appear to have some degree of functional unity.

Four procedural matters or assumptions need to be considered in applying latent trait theory. The first is the assumption of a single dimension. There are multi-dimensional methods of latent trait analysis, but the models most likely to be used, particularly in job knowledge testing, are those that assume a single underlying trait. The assumption of unidimensionality does not pose a rigorous demand, but the items to be scaled should possess a reasonable degree of homogeneity. This is somewhat at odds with the more general statements in these reports and elsewhere that high levels of internal consistency are not absolutely essential in content-referenced tests. The argument has been that any score on a work sample test should represent some functional unity, with the elimination of any items that have 0 or negative correlations with the composite. For latent

trait analysis, however, absence of orthogonality is not enough; substantial levels of internal consistency are required. For a latent trait analysis, therefore, a factor or cluster analysis is needed so that independent scores can be obtained for each of the component dimensions. There will then be as many latent trait analyses as there are dimensions in the item pool.

The second assumption is the assumption of local independence. In general, local independence means that performance on one item is not dependent on performance on other items. The most obvious violation of the assumption occurs when there are contingent items; for example, if item 2 can be answered correctly only if item 1 has been answered correctly (as in certain arithmetical reasoning sequences and work samples), then the assumption of local independence has clearly been violated. Other violations of the assumption are less obvious, and there is no convenient test for violations. However, strong evidence of unidimensionality is generally accepted as evidence that the requirement of local independence has also been satisfied.

The third point is a procedural matter. The LOGIST computer program for the three-parameter model available from the Educational Testing Service (Wood & Lord, 1976) requires very large sample sizes, calling for 50 items and 2000 subjects as a desirable minimum for determining item characteristic curves. This requirement may, perhaps more than anything else, be responsible for the early lack of interest in latent trait theory among applied measurement specialists. New programs have made it possible to estimate item parameters reliably on substantially smaller samples. Bejar (1977), reported results of sex differences in item characteristic curves using 178 males and 143 females in independent analyses of 20-item scales. With the simpler Rasch model, Wright and Stone (1978) reported a study with

less than 20 items and 40 people. Despite uncertainty about the statistical power of estimates based on small numbers of people or of items, it seems clear that progress in computer programming is resulting in more efficient algorithms requiring less luxury in the size of either sample. Latent trait analysis is becoming practical.

Nearly all latent trait programs now in use are iterative programs. That is, they begin with the raw score as the first estimate of the underlying ability and then estimate item parameters. A second ability estimate is based on these parameters, and the whole process of estimation enters a second iteration. The procedure continues until the solution converges, that is, until successive iterations produce little or no change in the item and ability parameters. Some sets of data simply do not converge. Where this happens, it indicates that the data will not fit the model being used and, perhaps, an alternative model might be attempted. It is more likely, however, where convergence fails that it is less a problem with the model than a problem with the data fitting the assumptions of the model. The assumption of unidimensionality deserves particular attention. One or two aberrant items in an item pool that clearly do not fit the underlying dimension can be responsible for difficulties in achieving convergence. The problem may disappear when they are deleted.

The fourth point is the need to check the fit of the data to the model. The question of the fit of the model has two components. One is the degree to which the theoretical item characteristic curve will in fact fit the data. A normal ogive or logistic curve simply may not be acceptable fits for an item; checking this out is a simple curve fitting problem. The other component is the fit of people to the model. Some people may not respond as the model

indicates that they should. If the set of responses to the items on a test from a person of average ability follows the model reasonably well, he will get most of the easy items correct, will tend not to get so many intermediate items correct, and will not succeed in giving correct answers to most of the difficult items. If a person responds carelessly, however, there may be little in the proportion of correct responses across different difficulties. If the person is using a test-taking strategy that capitalizes in some unusual way on certain item characteristics, it may be easier for that person to get correct responses to some generally difficult items for which that strategy is especially useful than to get correct answers for easier items where the particular trick does not work.

Thus, both items and people may be identified which do not fit the model, and it may be necessary to delete aberrant items or people or both to achieve a satisfactory solution.

Other Latent Trait Models. Although most work in latent trait theory has been done with conventional paper-and-pencil test items, usually multiple-choice, that can be scored dichotomously, latent trait models can also be applied to observations of work sample performance or product. Some of the "items" in a work sample will, in fact, be dichotomously scored. An example is whether a mechanic disconnects battery cables in the right or wrong sequence. Other observations might be continuously scored, for example, the actual deviation of precise measurement from a specified measurement in drilling a hole. The latter item, of course, could also be scored dichotomously as either within or exceeding allowable tolerances.

By using a free response model (Samejima, 1973), one can incorporate into a single latent trait analysis all observations fitting

a particular dimension regardless of the scoring format. The advantage is that dichotomies are not necessary; a test developer can take special advantage of the added information included in continuous responses.

Although multidimensional latent trait models have been proposed, no work with any of them is known to the present writer. Nevertheless, if data are not unidimensional, one could certainly submit them to factor analysis or cluster analysis. The Bejar (1971) study applied the Samejima model to data both multidimensional and reported with continuous responses.

Scoring Procedures. Suppose that three-parameter item characteristic curves have been computed for each item of a test and that they differ substantially on all three parameters, particularly on the difficulty parameter. For low ability examinees, correct responses to difficult items occur primarily by chance. Their responses to these items are not very informative; in fact, information curves for items with high difficulty levels are very nearly at zero for low levels of ability, peaking with substantial levels of information only in the high ability ranges. Obviously, then, the traditional score of the number of items answered correctly is not as informative for low-ability people as the score obtained by ignoring (i.e., assigning a zero weight to) the very difficult items (Lord, 1968).

An optimum weight curve can be computed from the item characteristics curve to show the optimum weight to be assigned to each item, at each difficulty level, in scoring the test. Lord noted three facts about the optimum weight curve:

1. As ability increases, the optimum weight increases, the increase gradually becoming nearly horizontal, asymptotic to a value proportional to the discrimination parameter of the item characteristic curve. At high ability levels, then, optimum item weights depend on item characteristics, not on ability level.
2. Optimum weights are lower for lower levels of ability because of the increasing effect of random guessing.
3. At very low levels of ability, the optimum weights for difficult items become very close to zero.

There is a bothersome paradox in all of this: one must know the ability level one is trying to measure before one can determine optimal weights for measuring it! If scoring can be done by computer, a maximum likelihood estimate of the ability parameter can be estimated for each person (Lord, 1977). Alternatively, one can use, by computing the maximum likelihood estimator, the conventional raw score as a preliminary estimate of ability. This value can be used in charts of optimum weight curves to find a nearly optimal set of weights for a particular examinee. The test paper can then be scored to obtain a nearly optimal total score and, if needed, a conversion table can be established for converting these "nearly optimal" scores to standard score scales of ability. (Lord, 1977, offered a variation of this procedure.) The fact that the optimum weight of an item is independent of the weights of other items is what makes tailored testing practical; the score can be expressed along the same scale even if different items have been used in arriving at it. There are alternative scoring options. From the first fact noted above, it follows that the discrimination parameters of the item characteristic curves can be used as weights, without regard to examinee ability level, if measurement focuses primarily on the upper ability levels. A slightly different value can be derived for multiple-choice items where there is guessing:

$$w = \frac{a}{1 - c} \cdot \frac{p - c}{p}$$

where

- w is the weight to be assigned to an item,
- p is the conventional item statistic, the proportion giving the correct answer to that item,
- a is the discrimination parameter of the item characteristic curve, and
- c is the lower asymptote (the "guessing" parameter) of the item characteristic curve (Lord, 1977).

Scoring systems can also be based on item difficulty parameters. Considering items as if they were arranged in slightly increasing increments in the order of difficulty, and assuming that guessing does not occur, an individual's score on the test can be defined as the difficulty parameter of the next item just after the last item correctly answered. Another option is to use the average difficulty parameters of the items answered correctly; under certain conditions in tailored testing, it is essentially equivalent to the maximum likelihood estimator (Lord, 1974).

In certification testing, the importance of the choice of scoring method depends on the level of ability at which certification is to be granted. The choice of an optimal scoring procedure is critical if the decisions are to be based on minimal competencies which are relatively quite low. If competency implies a high level of ability, however, no great information loss occurs even if the simple number of right answers is used as the score. At these levels, the only advantage in using the parameters of the item characteristic curves as the basis for scoring is that they provide a standard scale independent

of the particular distribution of cases or particular set of items used in constructing it.

The standard scale is especially valuable in work sample testing. There are often missing data; either the observers disagree on what they have seen to the point where no consensus is possible, or they miss seeing something important, or the performance follows an atypical pattern. Latent trait scoring removes many of the problems these events pose for traditional scoring.

The above discussion has centered on the three-parameter model; scoring by the Rasch model is somewhat simpler, although it is also a maximum likelihood procedure calling for computer estimation for maximum precision. The only item parameter, of course, is the difficulty level. Scores may be expressed in terms of difficulty levels, as before, or in terms of units of measurement (Wright, 1977).

Advantages of Latent Trait Analysis. A latent trait analysis has as its principal advantage the fact that it offers genuinely formal measurement of a mental trait. That is, the principle of transitivity and the principle of additivity are both assured through the use of latent trait analysis. A second advantage in scoring tests by latent trait analysis is the increased precision in measurement. Since a work sample test is developed to make decisions about individuals, maximum precision in measurement is an ethical obligation.

For widespread operations, such as military qualifications testing, these facts (and its basic nature) make latent trait analysis independent of the idiosyncracies of time or place. That is, the score obtained by one candidate for promotion at one time in one location can be interpreted on the same scale, and with the same

precision, as the score of a different applicant obtained in a different physical location at a different period of time. It follows that promotional policies can be far more uniformly administered than through measurement techniques that place heavy reliance on samples of people for determining the scores.

An almost equally important advantage of the latent trait analysis, particularly for job knowledge tests, is that it provides an item pool which can be dipped into for a small set of items useful for measuring precisely the knowledge level of an individual candidate and a different set of items equally useful for a different candidate. Their scores will be on an underlying scale common to both examinees even with non-overlapping sets of items. For this reason, test security is less a problem with conventional job knowledge tests.

Contemporary society is greatly concerned about possible racial and sex bias in measurement; the latent trait analysis makes it possible to identify race-by-item interactions or sex-by-item interactions and to delete items contributing to such interactions from the item pool (Ironson, 1977). This is not to be confused with main effects due to race or sex, which may represent true differences between the groups, but it does identify where the probability of getting a correct response at a given ability level varies with race or sex.

Finally, latent trait analysis provides a firmer foundation for evaluations of possible adverse impact for minority groups or women. According to current Federal Executive Agency Guidelines (1977), a test is said to have adverse impact on a group when its rate of selection is less than 80% the selection rate of the subgroup with the highest percentage selected. The existing Guidelines of the Equal Employment Opportunity Commission (1970) require a test user to choose

the test with the lesser adverse impact when there are two or more possibilities of similar validity. It is not entirely certain that the level of validity is a governing issue; from the point of view of fair employment practices, there is a substantial social and governmental pressure to seek tests with lesser adverse impact, irrespective of their validities.

One response to this has been to look toward work sample testing as a possible approach to employee selection with lesser adverse impact than ordinary paper-and-pencil tests have. Schmidt, Greenthal, Berner, Hunter, & Seaton (1977) recently published a study in which they demonstrated less adverse impact for a job sample than for a written aptitude test.

A methodological question might limit the generalizability of their findings.* Figure 2 identifies two tests, each with a different test characteristic curve. The test characteristic curves show differences in the discrimination value and essentially equal difficulty levels. That is, the slope on Test 1 is steeper than that for Test 2. Assume true ability differences on the underlying latent trait indicated by A and B on the abscissa. If these true differences exist, they will be obscured if Test 2 is used and exaggerated if Test 1 is used; that is, relative to the true ability scale, the obtained differences between the two groups on Test 1 is greater than the true difference, while the obtained difference on Test 2 is less than the true difference. In either case, the obtained score difference distorts the true difference between the groups. A similar figure can

*The author, with Gail Ironson as senior author, is preparing a detailed critique of the article by Schmidt et al.; it is sufficient here simply to illustrate the nature of the argument.

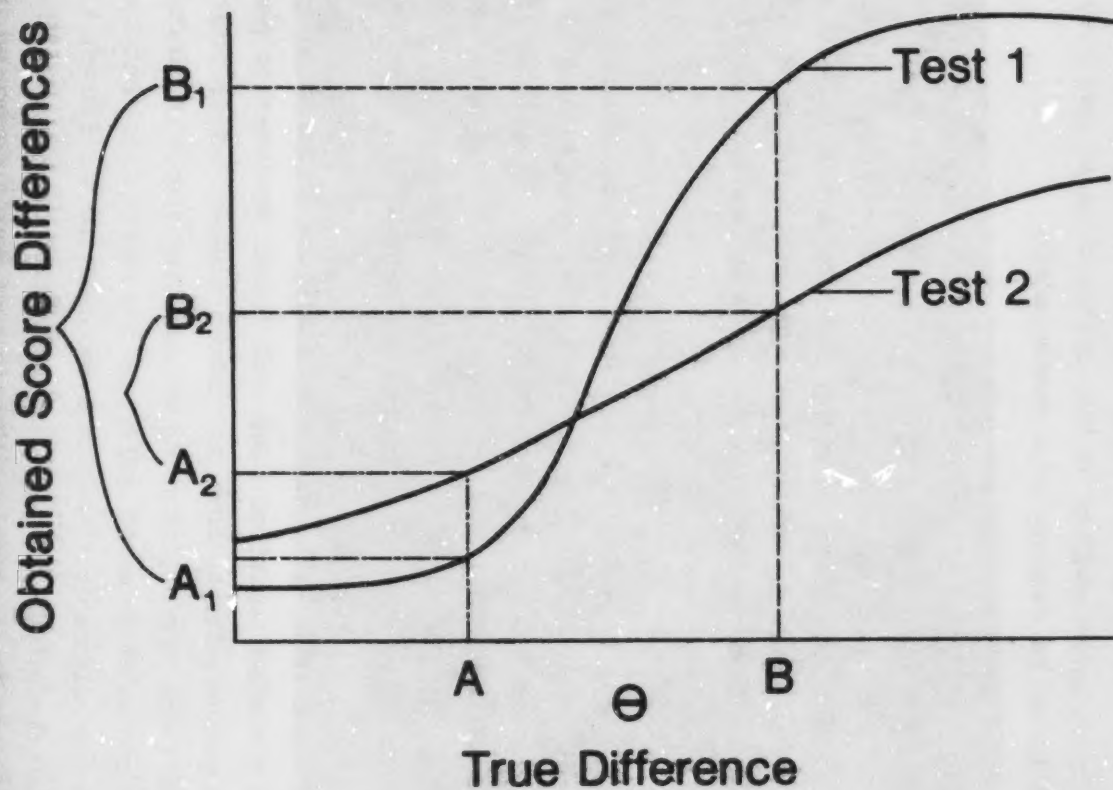


Figure 2. Obtained score differences between groups as a function of different test characteristic curve parameters

be drawn to indicate similar distortion in group differences in obtained scores where the difficulty levels vary and the discrimination values are essentially similar. Where the latent trait analysis of job knowledge tests and of any other tests with which it is to be compared, observed score differences can be interpreted in terms of the underlying latent ability so that different tests measuring the same ability can be compared on a common metric.

Possible Alternatives to Latent Trait Theory. The essential feature of latent trait theory is that it provides a content-referenced scale for interpreting scores and that it is not dependent on the distributional characteristics of specific samples. It provides a formal rather than the traditional psychometric form of measurement. The question arises whether there are alternatives to latent trait analysis to accomplish similar ends.

One possible alternative is the functional measurement advocated by Anderson (1976). It will not be discussed in this paper since it has not yet been applied to standardized tests. It is, however, a technique which develops a formal scale of measurement that is not dependent on the characteristics of individuals chosen for scale development; it is based upon theoretical mathematical functions.

One alternative is to interpret total test scores in terms of a score on a smaller subset of items with formal psychometric properties. The idea emerges, first, from the content standard scores proposed by Ebel (1962) and, second, from the writer's insistence on functional unity in content-referenced tests. It may be recalled that Ebel selected ten arithmetic items, each representing a different kind of arithmetic operation. No scale was identified for ordering these items. Scores on the total arithmetic test were interpreted in terms

of the number of items probably correct in this standard set of ten.

Suppose the ten items were scaled on some dimension. They might be scaled in terms of difficulty, the level of knowledge required, the social importance of the arithmetic skill involved, or some other dimension of interest. The scaling can be done using the method of equal-appearing intervals, and all items in the total test can be scaled. A small subset of items can be chosen to meet the requirements of a Guttman scale where the discriminial dispersions of the judgments do not overlap. The resulting items can be ordinally ranked, differences between items can be expressed in a common metric, and the metric can be expressed as the specific dimension of interest (difficulty, importance, etc.). The items form a content standard scale with known psychometric properties.

The subset of items can be used as a total test; however, it is generally believed, in keeping with the Spearman-Brown function, that a small set of items gives substantially less reliable scores than can be obtained from a larger set. Therefore, the conventional score on the total test from which the standard scale is drawn is a more reliable score. It can, however, be interpreted in terms of the standard scale by the simple expedient of developing a regression equation for standard scale values from total test scores.

The principle and procedures can be applied to any work sample where expert judges can form a reasonably reproducible scale along a dimension of interest. If the dimension of interest is difficulty, conventional item statistics can be computed as the scale value, and the "discriminal dispersion" can be expressed in terms of the conventional standard error, $1/\sqrt{pq}$. Brenner (1959) demonstrated that tests could be developed by these standard item analysis methods to yield satisfactory reproducibility coefficients.

CUTTING SCORES

Administrative convenience seems to demand that qualification testing be done with fixed standards of mastery designated by rigid scores, above which examinees are classified as masters, and below which they are classified as non-masters. There is no way to determine a cutting score empirically in the absence of assumptions or judgments. Shepard (1977) put it well: "Performance standards do not inhere in nature; they have to be decided upon by fallible people."

The use of a mastery cutoff point is, perhaps, inevitable, but the actual scoring of tests should be on a more refined scale. If the purpose of the cutting score is the classification of examinees, it should be recognized that the mere fact of misclassification is not the essential error. A serious error, perhaps the most essential error, is the degree of misclassification. Unless scores vary along a continuum, no evidence can be deduced for determining the degree of classification error. Certainly, if the only errors of classification are minor ones, the "close calls," e.g., the erroneous classification of people as masters when in fact they are almost masters, they are relatively minor. A truly serious error of classification is when an individual who is a true master, substantially above any minimum qualification, is mistakenly classified as a non-master, or in which someone who is wholly incompetent is mistakenly classed as a master.

Where attention to cutoffs must be given, the introduction to the topic by Hambleton et al. (1978) is worth noting:

"The problem of determining cutoff scores for assigning examinees to mastery states based on their criterion-referenced test performance has received much attention from researchers in recent years. Still, the problem seems far from resolved. The

arbitrariness of the proposed solutions has proved troubling to some measurement people, to the point where they seriously question the merits of determining and using cutoff scores at all (p. 26-27).

Despite the psychometric futility of cutting scores, there is no administrative procedure more widely accepted for test score interpretation than the establishment of some sort of cutoff score. Moreover, for some procedures of evaluating the usefulness of a test, standard reference points on the score distribution, similar to cutting scores in the way they are chosen, are necessary. The problem will not go away just because it is psychometrically intractable.

Four different ways can be used to establish such scores. All involve judgments, and each involves the collection of some kind of data as a basis for the judgment.

1. Normative determination. Although job knowledge and other work sample tests are generally intended to lead to content-referenced interpretations, cutoff scores may nevertheless be norm-referenced interpretations. One may determine a priori that some specified percentage of a specified sample of people can be assumed to be masters. These percentages are typically exercises in imagination, drawn from the air. Once the judgment has been made, however, the determination of a cutting score requires a distribution of scores from an appropriate sample of people.
2. Absolute decisions. It is sometimes argued that, if the job content domain has been properly sampled, and if the test content domain has added no irrelevancies, anyone who is really a master of the content domain should be able to pass all items in the tests. This means that the cutting score for mastery is set at the absolute value of 100%. Failing even one item results in being classified as a non-master. The arrogance in establishing such a cutting score is clear when one considers that even the most carefully devised examination is subject to some error of measurement. Whether one computes a standard error of measurement or a

standard error of estimated ability, or whether one simply arbitrarily makes an allowance for error, some departure from the 100% mark is allowable as a permissible degree of error. Even in these circumstances, however, the 100% score is the intended mastery score; the actual cutoff score is placed at something less than 100% to take into account computed or judged errors of measurement.

3. Decision theory. Decision theoretical models have been established for cutting scores. These can be illustrated with reference to Figure 3. If it can be assumed that Type 1 errors of classification are equal in importance and cost to Type 2 errors of classification, an optimal cutting score -- that is, one which minimizes the errors of classification -- is the point at which the distributions of scores of masters and the distribution of scores of non-masters intercept. Rorer, Hoffman, Laforge, and Hsieh (1966) and Rorer, Hoffman, and Hsieh (1966) have provided procedures for determining these cutoff scores with other loss functions.
4. Estimation of item difficulties. Ebel (1972) has proposed a knowledge estimation procedure for setting cutting scores. The procedure requires that the items be classified both in terms of importance level and difficulty level. Within each group of items so defined, knowledgeable judges determine or estimate the number of minimally qualified people who will get the items correct. The passing score is a percentage based on the average of these estimates. Essentially this same procedure was used by the Educational Testing Service in determining cutoff scores for certification of teachers in the State of South Carolina (Educational Testing Service, 1976); the legal success of the procedure was established by the fact that this usage was accepted by the courts up through the Supreme Court.

It should be emphasized again that none of these methods has special merit particularly from a measurement perspective. What is required is not a statistically or psychometrically defensible method, but rather a consensus among knowledgeable judges that a specific procedure, and the result of that procedure, is justified.

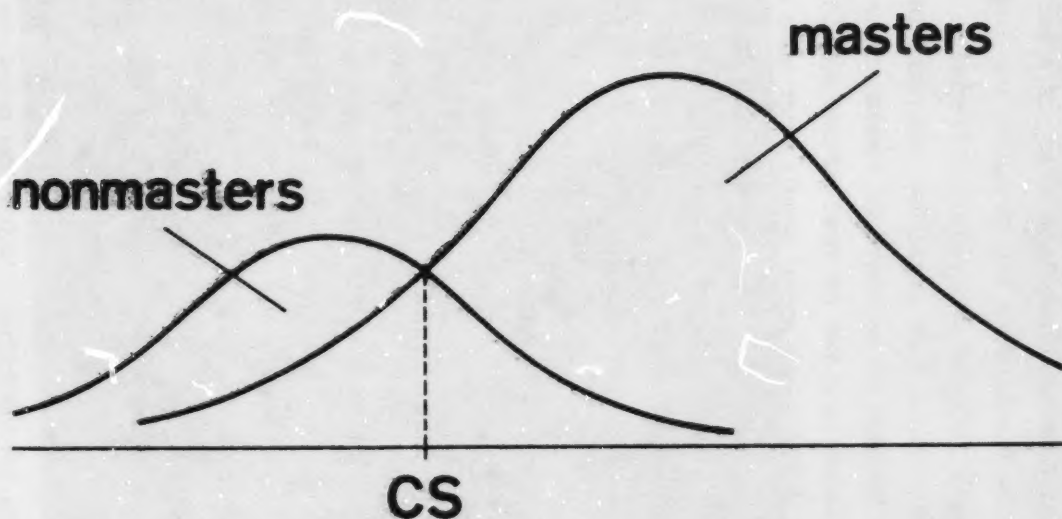


Figure 3. Cutting score or point of intersection of score distribution of masters and of nonmasters

GENERALIZABILITY ANALYSIS

Generalizability theory, as described by Cronbach et al. (1972), is an application of the principles of analysis of variance to estimate the magnitudes of specified sources of error in measurement. The simplest design for a generalizability study is a person-by-item design. That is, variance estimates are obtained for individual differences across people and for individual differences across items. An interaction term is also possible, but this is confounded with random error and becomes an estimate of random error variance. Much of the total variance in a set of test scores should be attributable to individual differences among people, but some of it is due to differences in samples of items. The generalizability analysis in this case identifies total variance as a proportion of variance due to individual differences among people, variance due to a main effect of items, and variance due to error.

More complex analyses can be designed for specific situations. In one common kind of situation, a group of people is given a work sample test in each of several different installations, in each of which there may be several different observers. In analysis of variance terms, persons are nested within observers who in turn are nested within installations. Complicating the situation is the frequent assumption that work sample observations made under conditions of probationary judgments generalize to a later time when the work sample is observed under less severe institutional control. A check on the assumption requires two conditions of measurement. The necessary research design can be described in Figure 4 with a Venn diagram identifying the potential sources of variance, overlapping circles indicating interactions. Using the notation of Brennan (1977), the design can be identified as $p:o:i \times c$, persons nested

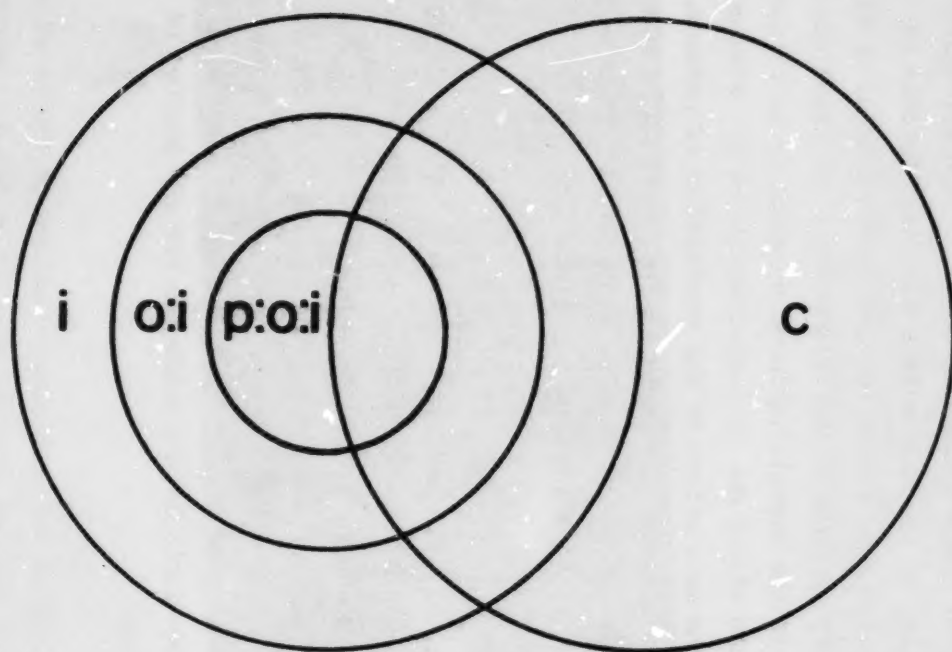


Figure 4. Venn diagram identifying variance estimates for persons (p) nested in observers (o) nested in installations, all crossed with conditions: c, i, o:i, p:o:i, i x c, o:i x c, and p:o:i x c

within observers nested within installations by condition. This experimental design makes possible seven estimates of variance sources, one for the total confounding of persons, observers, and installations, another for the confounding of observers and installations, another for the effect of installations alone. It will be recognized, at this point, that no main effect can be determined for persons or for observers independently of installations; however, a main effect for installations is possible; if the main effect for installations is essentially the same as the main effect for the confounding of observers and installations, it may be assumed that differences in installations contribute very little to the overall variance -- which is as it should be. However, if the main effect due to installations is not substantially different from the main effect of the confounding of persons, observers, and installations, then it would also follow that persons are representing very little contribution to total variance -- which is not the way it should be in attempting to measure individual differences.

Continuing the list of contributions to variance, the study will investigate a main effect due to condition and three interaction terms, one for each of the three levels of confounding of persons, observers, and installations by condition. The interaction terms, in each case, may be taken as error terms and, when one is interested in measuring the proficiency of persons, effects due to condition, installation, or observer-installation confounds are all sources of systematic error variance.

The principal usefulness for work sample testing of the paradigm is that it identifies the sources of error and, in so doing, identifies the limits to which scores on a test may be generalized. Suppose, for example, that there is indeed a strong main effect due either to

installations or to conditions. It would follow then that scores obtained in one installation under one condition would not generalize to other conditions. It would be possible, using the Cronbach, et al. equations, to develop an estimated observed score for specific combinations of installations and conditions. Insofar as the interaction terms are relatively small, therefore, a test may be said to be useful even across circumstances which have in fact a substantial main effect. If, however, the interaction terms are substantial, the error corrections are of questionable reliability and the test lacks dependability in measurement.

A generalizability analysis could be used to check on such effects as racial bias by using ethnic identification as one of the facets or conditions of measurement. In military applications, a work sample test may be given under highly standardized conditions that are somewhat aseptic. This is desirable from the point of view of precision in measurement, but it poses severe problems from the point of view of the generalizability of the results of that measurement. Does it necessarily follow that people who do well under the aseptic conditions of, for example, a training post will also do as well under the more realistic conditions of performing the same tasks in the field, in the rain, under conditions of jungle heat or arctic cold, or perhaps under combat conditions? While it is highly unlikely that an experimental study will include actual combat conditions, it is logical to design a generalizability study to facet a series of conditions ranging from highly facilitating to highly hostile. The REALTRAIN program (Shriver et al., 1975) offers an example of conditions which are not aseptic but which represent a closer approximation to combat than would be true in, let us say, a training center. Taking the same kinds of observations in both kinds of situations produces an opportunity to determine the degree to which variations in

the hostility of conditions account for variance in the overall performance of individuals.

The topic of generalizability is of such great importance that it is singled out for the fourth report of this series.

EVALUATIONS OF WORK SAMPLES

JOB RELEVANCE

The job relevance of a work sample test, particularly if it is a direct work sample, is almost never seriously questioned. The relevance of a job knowledge test, or even of the most highly abstracted work sample, is usually assured if efforts to assure have permeated all phases of test development from job analysis through the choice of scoring procedures. This is the singular beauty of work sample testing. Of the six requirements for a test to be accepted as an operational definition of a variable (Guion, 1977, in press), only two points of question seem likely to arise.

The first of these is the basic question of the suitability of the test content universe and domain relative to the job content domain. If the judgment is that knowledge or abstractable skill is to be measured, especially if measured by paper and pencil rather than hands-on methods, the test loses its face validity. It may be challenged as being designed more for the convenience of the tester than to test empirically justifiable prerequisites to effective performance. Claims of job relevance are more secure if judgments at this step are considered and explained and the process, reasoning, and degree of consensus are documented extremely carefully.

The other point of possible contention is in the scoring. Scoring procedures for a typical job knowledge test, whether traditional or utilizing latent trait analysis, are likely to be quite straightforward and, of themselves, be subject to little challenge. Such scores may, however, be contaminated by irrelevant variables such as reading ability or specific forms of test wiseness. Plausible hypotheses of contaminants in direct work sample may change the relevance of these tests, too. To the extent that such questions seem reasonable, studies evaluating the construct validity of the test or of the test inferences from the scores may be necessary.

RELIABILITY

Conventional reliability estimates are as applicable to work sample tests as to others. One may compute coefficients of stability or of internal consistency -- remembering, in the latter case, that the internal consistency of content-referenced tests need not be particularly high so long as there is some. Since the recommended procedures have been to include a large enough item pool to permit the generation of roughly parallel forms, even something similar to the classical coefficient of equivalence can be computed although the correlation of independently constructed domain samples is not a conventional estimate of reliability since they are not strictly parallel.

Reliability of Mastery Classification. There are also some special problems in estimating reliability for content-referenced tests. Many special methods have been proposed; many of them really yield estimates of the reliability of mastery classification schemes rather than estimates of the reliability of measurement (Livingston, 1976). The kappa statistic for correlating nominal data has been

used, with modifications, by Huynh (1976) and by Swaminathan, Hambleton, & Algina (1974) for determining an estimate of percentage of agreement in classification that takes into account the agreement that might be expected solely by chance.

It should be recognized, of course, that the reliability of classification depends on the placement of the cutting score; Huynh has pointed out a non-linear relationship between kappa and the cutting score. As the cutoff score gets larger, kappa increases up to some maximum, after which it decreases. Test length and test variability also influence the kappa statistic. Subkoviak (1976) presented a different method of estimating the reliability of classifications, a coefficient of agreement defined as the probability that an individual will be assigned to the same mastery state on parallel tests. The definition does not require a limitation of only two mastery states.

Reliability of Measures. The reliability of the scores as measurement, rather than as a basis for decision, has been approached by different authors in somewhat different ways, but commonly involving discrepancies between the obtained score and the point on the score distribution defined as the standard or cutting score. The analogy to norm-referenced testing is straightforward; in conventional classical reliability estimation, one is basing the estimate on the discrepancies between obtained scores of individuals and the mean of the distribution. Many of these techniques are essentially equivalent to classical reliability if the cutting score happens to be at the mean.

Livingston (1972) developed an equation for computing the reliability of measurement using the equation

$$r_{cc} = \frac{r_{xx} S_x^2 + (\bar{X} - C)^2}{S_x^2 + (\bar{X} - C)^2}$$

where

r_{cc} is the content-referenced correction of r_{xx} ,

r_{xx} is the classical reliability coefficient,

S_x^2 is the classical test variance,

$\bar{X}$ is the test mean, and

C is the standard or cutting score.

More recently, he has extended the basic idea to provide a statement of the reliability of a single score (Livingston, 1976). This modifies the above equation by setting r_{xx} at zero (to characterize a joint probability distribution of variation due to irrelevant conditions) and by replacing the mean of a distribution with an estimate of the individual's true score. The resulting equation is

$$r_{ss} = \frac{(T - C)^2}{S_e^2 + (T - C)^2}$$

where

r_{ss} is the reliability of the single score of one examinee,

S_e^2 is the error variance, the square of the standard error of measurement,

T is the examinee's estimated true score, and

C is the standard or cutting score.

The reliability estimate in this case is, therefore, somewhat analogous to a statistical test of significance in that the reliability depends

on how far an individual's true score is from the criterion score. If the true score and criterion score are identical, the reliability in the single score is given as zero.

The estimate of the reliability of content-referenced measures recommended by the present author has been developed by Brennan and Kane (1977) and combines the Livingston approach and generalizability theory. Brennan and Kane pointed out that the $\bar{X}-C$ term is fundamentally concerned only with errors due to sampling people; if one is also concerned with errors due to sampling items, as implied by the entire notion of domain sampling, a somewhat more complex determination is needed. To be consistent with generalizability theory, they prefer to call the measure they have proposed an index of dependability rather than an estimate of reliability.

The classical assumption in psychometric theory is that an obtained score equals a true score plus an error score. Brennan and Kane start from a more complex linear model:

$$X_{pi} = \mu + \pi_p + \beta_i + \pi\beta_{pi} + e_{o(pi)}$$

where

μ is the grand mean in the population of persons and universe of items,

π_p is the effect for person p ,

β_i is the effect for item i ,

$\pi\beta_{pi}$ is the effect for the interaction of person p and item i , and

$e_{o(pi)}$ is the error with o representing a replication subscript

Since the ordinary case of affairs provides only one observation for each person-item combination, the error term and the interaction term

are totally confounded and can for practical purposes be combined.

From this initial assumption that an obtained score is a deviation from a grand mean of possible scores according to the influence of the characteristics of the person being measured and the characteristics of the items chosen for measuring him, plus the ubiquitous error, Brennan and Kane derive an equation for estimating the index of dependability, noted here as I_d .

If the items are scored along a continuous scale, an analysis of variance procedure can be used to estimate I_d . Where the items are scored dichotomously, and the estimation can be simply expressed as

$$I_d = 1 - \frac{1}{n_i - 1} \left[\frac{\bar{X}_{PI}(1 - \bar{X}_{PI}) - S_{pi}^2}{(\bar{X}_{PI} - C)^2 + S_{pi}^2} \right]$$

where

I_d is the index of dependability,

n_i is the number of items,

$\bar{X}_{PI}$ is the grand mean of the item scores (which are either 0 or 1) over all persons and all items,

C is the standard or cutting score, and

S_{pi}^2 is the variance of the mean scores of persons over items.

The essential and important feature of this approach is its consideration of errors of sampling people and of sampling items. For content-referenced measurement, this double consideration is very important. The final test administered to an individual consists of only a sample of the items that might have been developed from a domain, a fact with special importance when scores are to be interpreted with reference to that domain.

A Coefficient of Accuracy. Before leaving the topic of reliability, an interesting new statistic specifically designed for content-referenced testing is the coefficient of accuracy proposed by Shaycoft (1977). It is not a reliability coefficient but is an analog of reliability. Using the example of a clinical thermometer with a systematic 2° error, Shaycoft argued that the thermometer would not be satisfactory for its purpose even if the readings it yielded were perfectly reliable. By analogy, a content-referenced test containing systematic error in the scores might yield a high reliability coefficient because of the systematic error. These observations suggest a need for a new psychometric statement which she called a coefficient of accuracy, analogous to a reliability coefficient except that it is reduced by any measurement error, be it random or systematic.

She also proposed an accuracy analog for the standard error of measurement, and provided a basis for correcting the coefficient of accuracy for variations in range. With no experience using the coefficient of accuracy, the writer is unprepared to make recommendations about it other than to suggest that its implications be studied; the concept has potentially great practical significance for content-referenced testing.

Modern Replacements for Reliability. Particularly in direct work sample tests, the errors due to differences among observers may account for a substantial portion of the error variance in overall scores. Moreover, work sample testing is usually conducted primarily for the purpose of identifying maximum potentiality; for this reason, some variance might be due to differences in motivational arousal, such as differences between conditions of institutional control, which tend to maximize motivation to perform, and conditions of field observation, in which the motivation to perform may be not quite so high.

Neither kind of differences, however, fit under the usual rubric of reliability estimation; they are much more appropriately described under the generalizability rubric.

The generalizability of measurement is an essential characteristic in work sample testing. Work sample observations which are highly dependent upon the place, time, or condition of observation, or on who does the observing, are less useful than those that can be depended upon to give common results under different conditions. It is argued, therefore, that generalizability, as the more general case in reliability estimation, is the critical consideration in the evaluation of work sample testing.

Latent trait analysis produces a precision estimate that also supersedes conventional reliability, the information function. Latent trait analysis does not preclude the use of a generalizability study. The primary value of latent trait analysis is as a scaling procedure; generalizability analysis investigates the sources of error in obtained scores, regardless of the metric used for describing those scores. Generalizability theory estimates the magnitude of various errors; latent trait theory minimizes error.

A content-referenced test, constructed and scored according to the principles of latent trait analysis, can be shown from the test information curve to have specific limits of probability error at different score levels. This is an estimate of the degree of precision in measurement; the precision of the estimate itself cannot be matched by any of the conventional techniques of reliability estimation.

VALIDITY

Notions of validity seem almost superfluous in discussions of work samples, at least of "total job" or "direct" work samples. For many such tests, only so-called content validity, described in these reports as job relatedness, seems important.

The evaluation ordinarily meant by the term content validity has already been described here as job relevance. The essential evidence of the validity of scores on a content sample stems not from statistical analysis of the scores themselves but from an evaluative analysis of the judgmental processes involved in developing the test. As Messick (1975) has pointed out, what has been called content validity is better described as content-oriented test development.

The panels of judges who evaluate items in the item analysis are at the same time providing the fuel for the evaluations of the test as a whole. The job relatedness ratio for items, which Lawshe (1975) called a content validity ratio, can be turned into a job relatedness index, which is simply the mean of the job relatedness ratio values of the items in the final form. It is an index number, expressed on a scale from -1 to +1, describing the degree to which knowledgeable expert judges perceive overlap between the ability to do a job and the ability to answer the items of the test correctly.

Construct Validity. Part of what people mean by content validity is a special case of construct validity. Therefore, disconfirmatory studies need to be conducted to investigate the possibility that variance in a set of test scores is attributable to characteristics other than the knowledge or proficiency inferred from the scores. For a job knowledge test, this is probably not a terribly serious

problem. However, as pointed out in the section on job relatedness, one might encounter the criticism of a given job knowledge test that scores on the test are as influenced by ability to read as by knowledge of the job. If the job content domain includes a great deal of reading, such a criticism is not likely to be seriously proposed. If, however, the job is one in which little or no reading is involved, then a job knowledge test consisting of multiple-choice items is likely to be considered contaminated by including a component not in the job content domain and therefore an irrelevant source of variance. The logic of construct validity needs to be invoked in evaluating this possible interpretation of obtained scores.

It might be done in either of two ways. One might determine the readability level of the test and decide whether it is high enough to reduce the scores of people who are otherwise qualified to take the job. Or, scores on the test can be correlated with scores on a standard measure of reading proficiency. If the correlation is high, the alternative interpretation of scores on the job knowledge test is supported. If it is low, however, it indicates that only a small proportion of the total variance in the test scores can be attributed to reading ability.

A common problem in content-referenced measurement is that an available sample is very likely to have low variance among scores; construct validity studies based on correlational research will yield low correlations in such samples, failing to show relationships that may actually exist. For this reason, among others, it has been argued that content-referenced tests should be developed to yield a reasonable spread of scores.

Consistency in Domain Sampling. In discussing what was called

content validity, Cronbach (1971) suggested the independent construction of two tests from the same content domain and test specifications. The recommendations for test development in this report called for developing substantially more than twice as many items as might be needed. This makes it possible to conduct some item analysis research, identify and discard poor items, and still have a sufficiently large item pool to permit the allocation of items into two forms. (This is not, of course, precisely the operation prescribed by Cronbach, but it does provide similar samples of the domain.) The validity of the sampling procedure can be assessed by correlating the total scores on these two independent sets of items. More important is that the forms provide an important facet in generalizability studies.

Predictive Utility. While there is no great objection to doing predictive validity studies to determine whether the job knowledge test does in fact predict future performance, neither is there any particularly good reason for doing so. While a high predictive validity coefficient between scores on a work sample and some later measure of performance suggests additional evidence of the construct validity of inferring proficiency from work sample scores, a low predictive validity coefficient would not cast doubt on the validity of such inferences. Rather, it would cast doubt on the validity of inferences from the criterion measure; unless the criterion is another work sample, or the same work sample observed under other conditions (in which case we are discussing a generalizability model), the criterion is unlikely to be as carefully constructed or as job related as the work sample itself. It is true that the logical foundation for using a job knowledge test for placement decisions is an implied prediction. Nevertheless, the implied prediction is simply that, if a high scoring person is placed on a job, he will be able to do the job immediately.

Criterion-related validation can be useful in evaluating abstracted work samples or job knowledge tests. Although abstractions, such tests are intended to be used for inferring proficiency. Beginning with the definition of a job content universe to the definition of a test content domain, the assumption and development of a job knowledge test or other abstraction has consistently been that the knowledge or skill being tested is an essential prerequisite to successful performance on the job. This is a hypothesis; it is a hypothesis that performance on one variable, such as job knowledge, is related to performance on a different variable, proficiency. This is a correlational hypothesis, and it should be tested following the principles of criterion-related validation. Such a study may not be necessary if there is sufficient agreement among qualified judges that the knowledge tested is indeed prerequisite to effective performance.

SUMMARY: PRINCIPLES OF WORK SAMPLE TESTING

This paper has concentrated on the application of general principles of measurement theory and on general principles of the evaluation of psychological measurement to the construction and evaluation of work sample measures of performance proficiency. It started with an emphasis on job analysis as the foundation for work sample test development, and it continued with idealized suggestions for test construction and evaluation for various degrees of work sample abstraction, that studies may be needed to demonstrate that performance on an abstract work sample is consistent with performance on the job itself. The greater the degree of abstraction, the greater the necessity for empirical verification of that relationship.

In summary, seven principles can be drawn from this report which should be observed in the evaluation of work sample tests:

1. The choice of a job content domain needs to be justified. The job content domain is a portion of a total job content universe. It is chosen on the basis of expert judgment, not because it is representative of the universe, but because it is an important or salient aspect of the universe for the purposes of the decisions that are to be based on the testing. In a situation in which some aspects of the job content must be learned after one is placed on the job (Gael, 1977), a selection or certification test should not include those portions of the job content universe. Rather, it should be restricted to a domain of job skills or knowledge or activities that a candidate for the job is expected to bring to it. Other kinds of purposes may impose other kinds of restrictions. For skill qualification testing, the restrictions should probably be minimal; that is, to certify competence to perform a job, the job content domain should probably be close to the total job content universe. It might still, however, represent a substantial amount of abstraction from that universe to eliminate redundancies, or to emphasize the most important aspects of the universe, or for other reasons. Any condensation of the job content universe into a smaller job content domain should be accompanied by careful documentation of the reasoning used, and it should also record the reliability of the independent judgments of panel members and the degree of consensus achieved through panel discussions.
2. The test content domain and the job content domain should be as congruent as possible. Measurement of that congruence is a matter of judgment. Numerical indices like Cronbach's correlation of independently developed content samples, the Lawshe content validity index, or the Rovinelli-Hamblen index of test-objective congruence may all be used, but none of them should be taken too seriously. For one thing, the two domains are defined in abstract terms and any actual attempt to make them literally measurable would probably distort them. For the other, the important evidence consists of the judgments of qualified experts, not of statistical indices. Such indices are valuable, but they are valuable precisely because they provide a means by which expert judgments may be systematically collected and analyzed.
3. Scoring procedures should approximate formal, fundamental measurement as much as possible. Special care should be taken to assure that scores are indeed established along a transitive scale of measurement. In practice, any set of

numbers ultimately assigned is, of course, mathematically transitive; that is, if individual a gets 10 points, b gets 15 points, and c gets 20 points, then obviously b has a higher score than a, c has a higher score than b and also has a higher score than a. If, however, the 10, the 15, or 20 points earned are earned on quite different bases, it does not follow that b is necessarily more proficient than a, or that c is more proficient than b, or that c is more proficient than a, even if the other two statements were true. Obviously, the call for transitivity is a call for reasonable homogeneity or functional unity in the system of scoring. If this unity cannot be achieved, individual subsets of scores should be obtained; the importance of having such subsets can be inferred from their intercorrelations. It is better to try to develop the subsets and find out they are not necessary because of high intercorrelation than simply to assume that the overall score is sufficiently homogeneous.

4. Levels of proficiency should be measured: scores should not be merely dichotomies. Test construction, particularly following the recommendations of this report, is too expensive in time and resources to allow the deliberate loss of information caused by dividing an entire distribution of scores into just two parts. By using continuous scoring, a test can continue to be used even when standards change (as for different levels of a common career ladder or in response to new circumstances). Moreover, most of the appropriate procedures for evaluating the test — reliabilities, information functions, construct validities, generalizability coefficients — require variance along a continuous scale of measurement.
5. The opportunities for irrelevant influences on individual scores should be at a minimum. Insofar as all testing procedures are carefully standardized, and insofar as there is some variability in performance, this principle is essentially an admonition to check for violations of the assumption of construct validity. It is better, however, to recognize that work sample tests may, in a well prepared sample of subjects, result in very low variances, in which case investigations into alternative construct explanations are difficult. Even in these cases, however, it is very important to make the effort to ascertain the possible influence of attributes of observers, attributes of conditions, or simply irrelevant attributes of the people being measured on the performance of the work sample.

6. Work sample scoring, if the test is to be used in an organization with diverse locations, should be standardized on a scale of reference that is applicable to an organization-wide population. Some form of content standard score is preferable to a standard normative scale of measurement. While normative interpretations may not be irrelevant for many uses, the principal meaning of a set of work sample scores should be inherent in the content of the test. Number or proportion of items answered correctly, number of points awarded on an observer scoring form, or simply a dichotomous pass-fail notation may be seen as roughly a content standard score. Nevertheless, more sophisticated procedures can and usually should be used, such as latent trait scaling, or keying scores to special Guttman scales, or functional scaling.
7. Scores from work sample testing in the usual conditions of institutional control must generalize not only to field settings but to a variety of field settings. Work sample testing, before it is made operational on a large scale, should be subjected to appropriate generalizability analysis.

The recommendations of this report for work sample testing are extensive; following all of them for very many jobs would be prohibitively expensive and laborious. The use of a panel of experts from the beginnings of job analysis to the completion of test development would come to countless manhours of deliberation. The recommended calibration by latent structure analysis and evaluation by generalizability studies could cost many thousands of dollars for data collection alone. The expenditures of time and money may be justified for certain extremely critical jobs, but not for many. Moreover, the world of measurement does not contain enough test development specialists to conduct such intensive campaigns for effective work sample measurement for more than a few very important job categories.

The conclusion is inescapable that short cuts are needed. If a work sample were developed according to all of the recommendations in this report, its job relevance and (where it matters) its validity

would be unquestionable. But will this be true if short cuts are taken? It probably will be with some short cuts, but we don't know which ones are "safe." The report has spoken repeatedly of abstracting from universes and domains; the job of developing work sample tests has a total job content universe as does any other, and short cuts represent similar abstracting. What short cuts or abstractions will yield work samples, direct or abstract, as unquestionable in relevance and validity as those developed without cutting corners?

It is an empirical, not a rhetorical, question. Systematic studies are needed to compare work samples developed by simpler procedures to model work samples developed by the more elaborate procedures outlined in this report.

The most important questions, however, are not questions of procedural simplification. They are questions of the optimal or permissible kinds and levels of abstraction from a job or test content domain in the development of abstract work samples or job knowledge tests. It is a problem in generalizability. Will abstractions of one, developed by one set of rules, generalize more or less well than abstractions of a different kind, or developed by a different set of rules, to model work samples developed with very little abstracting at all?

REFERENCES

- Anderson, N. H. How functional measurement can yield validated interval scales of mental qualities. Journal of Applied Psychology, 1976, 61, 677-692.
- Asher, J. J., & Sciarrino, J. A. Realistic work sample tests; a review. Personnel Psychology, 1974, 27, 519-533.
- Bejar, I. I. An application of the continuous response level model to personality measurement. Applied Psychological Measurement, 1977, 1, 509-521.
- Boyd, J. L., Jr., & Shimberg, B. Handbook of performance testing. Princeton, N.J.: Educational Testing Service, 1971.
- Brennan, R. L. Generalizability analysis: Principles and procedures. (ATC Technical Bulletin No. 26). Iowa City: American College Testing Program, 1977.
- Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.
- Brenner, M. H. An experimental comparison of cluster-analysis-refined spatial relations items constructed by two methods: Guttman scale analysis and item analysis. Unpublished masters thesis, Bowling Green State University, 1959.
- Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. An investigation of sources of bias in the prediction of job performance: A six-year study. Project Report PR-73-37. Princeton, N.J.: Educational Testing Service, 1973.
- Comer, J. C. Trade competency testing via specimen inspection. Journal of Industrial Teacher Education, 1971, 9, 50-54.
- Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.) Educational measurement (2nd ed.) Washington, D.C.: American Council on Education, 1971.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.
- Duffy, P. J. Development of a performance appraisal method based on the duty module concept. (Technical Paper 273). Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences. 1976.

- Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.
- Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1972.
- Educational Testing Service. Report on a study of the use of the National Teachers Examinations by the state of South Carolina. Princeton, N.J.: Educational Testing Service, 1976.
- Equal Employment Opportunity Commission. Guidelines on employee selection procedures. Federal Register, August 1, 1970, (No. 149), 12333-12336.
- Federal Executive Agency Guidelines. Uniform guidelines on employee selection procedures. Federal Register, December 30, 1977, (No. 251), 65542-65552.
- Foley, J. P., Jr. Overview of Advanced Systems Division Criterion Research (maintenance). Paper presented at Symposium on Criterion Development for Job Performance Evaluation, San Antonio, Texas, 23-24 June, 1977.
- Gael, S. Development of job task inventories and their use in job analysis research. Catalog of Selected Documents in Psychology, 1977, 7(1), 25. (Ms. No. 1445).
- Grant, D. L., & Bray, D. W. Validation of employment tests for telephone company installation and repair occupations. Journal of Applied Psychology, 1970, 54, 7-14.
- Guilford, J. P. Printed classification tests (Report No. 5). Washington, D.C.: Government Printing Office, 1947.
- Guion, R. M. Content validity — the source of my discontent. Applied Psychological Measurement, 1977, 1, 1-10.
- Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.
- Hively, W., II, Patterson, H. L., & Page, S. H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

- Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.
- Ironson, G. H. A comparative study of several methods of assessing item bias. Unpublished doctoral dissertation, University of Wisconsin-Madison, 1977.
- Jones, A., & Whittaker, P. Testing industrial skills. New York: Wiley, 1975.
- Lawshe, C. H. A quantitative approach to content validity. Personnel Psychology, 1975, 28, 563-575.
- Livingston, S. A. A criterion-referenced application of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.
- Livingston, S. A. The criterion-referenced reliability of a single score. (COPA 76-01). Princeton, N.J.: Educational Testing Service, 1976.
- Lord, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology, Vol. II, San Francisco: Freeman, 1974.
- Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.
- Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.
- Maier, M. H., Young, D. L., & Hirshfield, S. F. Implementing the skill qualification testing program. (R & D Utilization Report 76-1). Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1976.
- Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.
- Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.
- NCARB. NCARB Architectural Registration Handbook. Washington, D.C.: National Council of Architectural Registration Boards and Architectural Record Books, 1976.

- Peterson, D. A., & Wallace, S. R. Validation and revision of a test in use. Journal of Applied Psychology, 1966, 50, 13-17.
- Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.
- Root, R. T., Epstein, K. I., Steinheiser, F. H., Hayes, J. F., Wood, S. E., Salzen, R. H., Burgess, G. G., Mirabella, A., Erwin, D. E., & Johnson, E., III. Initial validation of REALTRAIN with Army combat units in Europe. (Research Report 1191). Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1976.
- Rorer, L. G., Hoffman, P. J., & Hsieh, K. Utilities as base rate multipliers in the determination of optimum cutting scores for the discrimination of groups of unequal size and variance. Journal of Applied Psychology, 1966, 50, 364-368.
- Rorer, L. G., Hoffman, P. J., LaForge, G. E., & Hsieh, K. Optimum cutting scores to discriminate groups of unequal size and variance. Journal of Applied Psychology, 1966, 50, 153-164.
- Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Tijdschrift voor Onderwijsresearch, 1977, 2, 49-60.
- Rubinsky, S., & Smith, N. Safety training by accident simulation. Journal of Applied Psychology, 1973, 57, 68-73.
- Rudner, L. M. Item and format bias and appropriateness. Washington, D.C.: Model Secondary School for the Deaf, 1976.
- Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38, 203-219.
- Schmidt, F. L., Greenthal, A. L., Berner, J. G., Hunter, J. E., & Seaton, F. W. Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and examinee attributes. Personnel Psychology, 1977, 30, 187-197.
- Schrivver, E. L., Mathers, B. L., Griffin, G. R., Jones, D. R., Word, L. E., Root, R. T., & Hayes, J. F. REALTRAIN: A new method for tactical training of small units. (Technical Report S-4). Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1975.

- Shaycoft, M. F. The coefficient of accuracy: A new statistic for criterion-referenced tests. Palo Alto, Calif.: American Institutes for Research, 1977.
- Shepard, L. Reporting results of competency-based measurement. Paper presented at meeting of National Council on Measurement in Education, New York, 1977.
- Shimberg, B., Esser, B. F., & Kruger, D. H. Occupational licensing: Practices and policies. Washington, D.C.: Public Affairs Press, 1973.
- Slinde, J. A., & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.
- Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.
- Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.
- Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.
- Thurstone, L. L. Primary mental abilities. Psychometric Monograph No. 1, 1938.
- Tryon, R. C., & Bailey, D. E. Cluster Analysis. New York: McGraw-Hill, 1970.
- Uhlaner, J. E., Drucker, A. J., & Camm, W. B. Army research in the criterion area: An overview. Paper presented at Symposium on Criterion Development for Job Performance Evaluation, San Antonio, Texas, 23-24 June, 1977.
- Wesman, A. G. Writing the test item. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.
- Wheaton, G. R., Fingerman, P. W., & Boycan, G. G. Development of a model tank gunnery test. (Technical Report TR-78-A24). Alexandria, Va. U.S. Army Research Institute for the Behavioral and Social Sciences, 1977.

- Wood, R. L., & Lord, F. M. A user's guide to LOGIST. Princeton, N.J.: Educational Testing Service, 1976.
- Woodson, N. I. C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64.
- Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: 1968.
- Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.
- Wright, B. D., & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model. (Research Memorandum No. 23A). Chicago: Statistical Laboratory, Department of Education, The University of Chicago, 1978.
- Wright, B. D., & Stone, M. H. Best test design: A handbook for Rasch measurement. Chicago: University of Chicago and Social Research, Inc., 1978.
- Yerkes, R. M. (Ed.), Memoirs of the National Academy of Sciences. Vol. XV, Psychological examining in the United States Army. Washington, D.C.: Government Printing Office, 1921.



ARI TECHNICAL REPORT
TR-79-A11

Principles of Work Sample Testing: IV. Generalizability

by

Robert M. Guion

and

Gail H. Ironson

BOWLING GREEN STATE UNIVERSITY

Bowling Green, Ohio 43403

April 1979

Contract DAHC 19-77-C-0007

Prepared for



U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Approved for public release; distribution unlimited.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR-79-All	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PRINCIPLES OF WORK SAMPLE TESTING: IV. GENERALIZABILITY	5. TYPE OF REPORT & PERIOD COVERED Final 15 Nov 1976 - 15 June 1978	
7. AUTHOR(s) Robert M. Guion Gail H. Ironson	6. PERFORMING ORG. REPORT NUMBER	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Bowling Green State University Bowling Green, Ohio 43403	8. CONTRACT OR GRANT NUMBER(s) DAHC19-77-C-0007	
11. CONTROLLING OFFICE NAME AND ADDRESS US Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, VA 22333	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q161102B74F	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	12. REPORT DATE April 1979	
	13. NUMBER OF PAGES 29	
	15. SECURITY CLASS. (of this report) Unclassified	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES Monitored by G. Gary Boycan, Engagement Simulation Technical Area, Army Research Institute		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Measurement theory, psychometrics, work sample testing, validity, content- referenced testing, criterion-referenced testing, generalizability theory		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Three kinds of generalizability are described: generalizability of test content, generalizability of relationships such as are described by validity coefficients, and generalizability of test scores across varying conditions. All are deemed important in the construction and evaluation of work sample tests. A special problem is determining the limits of generalizability; it is suggested that the research designs of generalizability theory can be applied in some cases to determining the limits of the generalizability of criterion-related validities.		

DD FORM 1 JAN 75 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

AUG 25 1980

230

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

20. (continued)

This report, the last of four, is written for psychologists and others interested in theories of testing.

231

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

PRINCIPLES OF WORK SAMPLE TESTING: IV. GENERALIZABILITY

BRIEF

Generalizability has been identified as a major concern in the evaluation of work sample tests. Generalizability is also an ambiguous term that has had substantially different meanings in different contexts. It sometimes refers to generalizability of test content to a broader test domain. It has often been used in the context of validity generalization, a special case of the generalizability of relationships. It has been recently used frequently to refer to generalizability theory, or an approach to the generalizability or dependability of scores.

An example of systematic effort to assure generalizability of content is cited from the development of a model tank gunnery test (Wheaton, Fingerman, & Boycan, 1977); this work is cited particularly for the development of an index of generalizability. Its applicability is particularly great for direct work samples. For abstracted work samples, special attention should also be given to the importance of specific task variables and to matching these variables in the construction of the abstracted sample.

Generalizability of relationships, as in the testing of a criterion-related hypothesis, is understood best in the example of the generalizability of criterion-related validities, but there are other examples. It is a criterion-related hypothesis that needs to be tested and, if supported, evaluated for its generalizability when a literal measurement of one variable is used for inferring measuring of a different variable; this is usually done under the rubric of construct validity. A similar example is the abstracted work sample, in which the literal performance of the unreal task is used for inferring performance on the real one. Studies of generalizability of relationships need to pay particular attention to the limits of such generalizability.

Generalizability theory, as proposed by Cronbach, Gleser, Nanda, and Rajaratnam (1972) refers to the generalizability of scores. Several designs for such studies are described. A general conclusion is reached that the limits of validity generalization (i.e., of the generalizability of relationships) might be studied by the generalizability theory research designs.

TABLE OF CONTENTS

INTRODUCTION.	1
GENERALIZABILITY OF CONTENT	3
A CASE STUDY IN ITEM SAMPLING	4
ABSTRACTED TESTS	6
INDUSTRY-WIDE RESEARCH	8
GENERALIZABILITY OF RELATIONSHIPS	9
THE CRITERION-RELATED HYPOTHESIS	9
Inferences from Literal Measures.	10
Abstracted Work Samples	11
VALIDITY GENERALIZATION.	12
The Schmidt and Hunter Study.	13
Limits of Generalizability.	14
GENERALIZABILITY OF SCORES.	15
SOURCES OF ERROR IN WORK SAMPLES	19
DESIGNS.	20
Dependability of Instructor Ratings	22
SUMMARY	27
REFERENCES.	29

LIST OF FIGURES

<u>Figure No.</u>		<u>Page</u>
1	Venn diagram identifying variance components for estimating generalizability of scores of persons.	21
2	Venn diagram for estimating variance components in student ratings of instructors. . . .	23
3	Two generalizability designs for estimating trainee and condition components	26

INTRODUCTION

Each of the preceding reports of this series has concluded that a major consideration in the evaluation of a work sample test is its generalizability. Unfortunately, the word is so ambiguous that the conclusion can mean different things to different readers. Consider the following list of statements drawn primarily from the other three reports:

1. In an experiment for evaluating programs or materiel, the sample of performance used as the dependent variable of the experiment should be representative of -- that is, should generalize to -- the performance in a target situation.
2. The properties of the tasks included in a work sample should be representative of the properties of the tasks in the domain from which they are sampled; for example, if, in the job content domain the job is structured so that each batch of 20 completed assemblies is a recorded unit, then the work sample should be based on a similar work unit. To the extent it does not, performance on the work sample may not generalize to performance on the real job.
3. The first step in the evaluation of a work sample is to evaluate the degree to which it is congruent with the domain being sampled; are the salient task characteristics adequately sampled in appropriate proportions?
4. The simple, abstracted work sample is best evaluated in terms of how well performance on the abstraction can be used to infer performance on a more highly complex real job. Performance on the abstract portion must generalize to performance on the whole.
5. In measurement in institutional settings, the actual measurements obtained usually are of less concern than are attributes to be inferred or predicted from them; the evaluation of measurement under conditions of institutional control is based on the degree to which the scores will generalize to attributes of greater institutional concern.

6. Measurement of one attribute is often inferred from literal measurement of quite a different attribute, such as the use of reaction time to infer some aspect of information processing. The usefulness of such an arrangement depends on how well performance on the one attribute generalizes to performance on the one inferred.
7. The criterion-related validity obtained in one setting may generalize to a different setting if the characteristics of the subject populations, tasks, and settings are essentially similar.
8. A disadvantage of tight, experimental control of measurement in a laboratory situation is that the situation may differ so substantially from the target situation that measurement in the one will not generalize to the other; where the situations differ markedly, "generalizability must be assessed."
9. The evaluation of work samples in field settings is incomplete unless it can be shown how well performance in the field setting used will generalize to performance in some targeted setting.
10. The variables that describe a task -- autonomy, demand for unwavering attention or other skills, amount of feedback, and others -- may change systematically as characteristics of the situations in which performance is measured change; as task variables change, generalizability may be reduced.

All of these statements are special instances of the broader statement that, regardless of purpose, setting, or combination of variables and methods of measurement, generalizability is the universally required characteristic of effective measurement. They do not, however, merely say the same thing in ten different ways. The first three statements refer to the generalizability of the content of the measurement; they deal broadly with the generalizability of the actual content of the content sample chosen to the rest of the domain that might have been chosen. The next four statements all deal with the generalizability of a relationship; they ask about the relationship of the variable measured to a variable one may perhaps wish had been measured, and, in statement 7, about the generalizability of such a relationship

across settings. The last three statements all refer explicitly to the generalizability of scores, or of inferences from scores, across different conditions of measurement.

In almost every one of the statements in that list, there is a sense in which all three kinds of generalizability may be relevant; nevertheless, these three can be examined independently both for the concepts they represent and for the possible methods available for evaluating them.

GENERALIZABILITY OF CONTENT

Every sampling is intended to generalize to the whole from which it is a sample. Random sampling in experimental work is intended to assure generalizability of results to the population sampled; stratified sampling is the effort of people whose faith in random sampling is small to assure such generalizability.

So it is with content sampling in the development of a work sample. Each stage in the process is a sampling exercise, not necessarily representative, which defines the part or aspect of the previous whole to which conclusions of some sort are expected to generalize. Job content domains are intended to generalize to certain aspects of job content universes; so it is also with test content domains and universes, and a test content domain is expected to have potential for generalizing to selected aspects of a job content domain. And the sample of the test content domain ultimately chosen is certainly expected to generalize to the test content domain; that is what so-called content validity is about. The problem is the choice of procedures by which one may feel comfortable in making generalizations from the sample to the whole.

A CASE STUDY IN ITEM SAMPLING

Wheaton, Fingerhman, and Boycan (1978) dealt explicitly with the generalizability problem in item sampling and developed an index of generalizability to be applied to "items" of a work sample. The work was that of a tank crew in "neutralizing" a target; the items were job objectives (of which they identified 266). Each job objective consisted of some subset of a possible 114 behavioral elements. By cluster analysis of a matrix of the 266 objectives and 114 elements, fourteen empirical and two rational clusters of objectives, or families, were identified. (In the third report of this series, the steps recommended for job analysis follow a similar pattern; to compare the two reports, job objectives can be equated with the component tasks of report III; behavioral elements to task elements; and families to task categories.)

The procedure followed was related by Wheaton et al. to the classical procedures of item analysis; they argued (a) that classical item analysis produced a set of items for a final test that would generalize to the whole and (b) that classical item analysis was too expensive to duplicate in a tank gunnery test. The present report will not follow the first of these arguments; what they did seems preferable to the more expensive item analysis. Empirical item analysis techniques will succeed in selecting a subset of all original items that will maximize the overall variance in that set. That is, a skillfully conducted item analysis, with certain sets of data, will result in the elimination of those items which have a low correlation with the composite of all other items, and the resulting test will have a variance comparable to that of any other similar-size subset of items and perhaps even a reliability as good as the total set.

Items eliminated, however, may be eliminated only because they are low in variance in the particular sample in which the item analysis is done. Alternatively, the items may be eliminated because they have low common variance, although substantial variance of their own, and therefore do not contribute to the internal consistency of the final test.

The argument of these reports is that high levels of internal consistency are not necessary in work sample tests, even though sub-scores on a work sample should have enough internal consistency to demonstrate some degree of functional unity. If this argument is accepted, it follows that an item need not be eliminated from the test (or subtest) unless it actually shows a zero or negative correlation with the rest of the test. Thus an empirical item analysis might have eliminated some items identified as important by the procedures used by Wheaton et al. -- items which helped assure that the final test was indeed representative of the behaviors actually required on the job.

The rational foundation for the approach is the assumption that the more behavioral elements that one job objective has in common with other job objectives, the more generalizable is the performance on it relative to the others. Arbitrarily but correctly, they believed that each family of job objectives should be represented in the final work sample according to the number of objectives it contained; that is, a family or cluster of objectives with 30 objectives in it should be represented three times as heavily as a ten-objective cluster. The issue was to choose the one or the three objectives as items of the work sample so as to maximize the generality of the resulting test.

Several methods were considered and rejected. For example, one

might take the objectives with the largest number of behavioral elements. This was rejected because many of the elements might occur only in that one objective, that is, they might not generalize to the family of objectives.

Two methods were considered intuitively sound and were ultimately combined. Each of them is a method of obtaining a weighted sum of the behavioral elements included in each job objective. One method of weighting assigns a weight equal to the frequency that element i occurs within a family, that is, F_i . This assigns the greatest weight to the most commonly occurring elements. The second method is to assign a weight of F_i/D_i , where D_i is the frequency with which the element occurs in the total domain; this method would give the greatest weight to an element which occurs almost exclusively within the one family, even though it may not appear very often. The two methods were combined by multiplying them so that the weight recommended by Wheaton et al. is $\Sigma(F_i^2/D_i)$, called the index of generalizability, for each job objective. Although their decisions were based in part on practical issues of cost as well as on these indices, a relatively inexpensive content domain could be sampled largely empirically through the use of such an index.

ABSTRACTED TESTS

The item sampling reported by Wheaton et al. is an example of the direct work sample; that is, the job objectives chosen were component parts of the actual jobs selected because they were more widely representative than others. In the abstracted work sample, the "items" -- or at least the composites which the set of items produce -- may be created rather than selected by sampling. The abstracted work sample identifies certain components -- task elements -- and puts them

together in such a way as to represent, if not to resemble, the essential or crucial elements of the job itself. Again, the purpose is to select a set of final tasks that will generalize to the whole of the original content domain, even if less intuitively obviously.

The task elements might be chosen by a procedure not unlike that for weighting elements in the Wheaton et al. example. An alternative is to have a panel of qualified judges (i.e., qualified because of their knowledge of the job) rate each element both for frequency and importance. Although these ratings will probably be highly correlated, they are both necessary if the task elements that occur rarely but are crucial when they occur are to be identified (or conversely, if there is to be any success in identifying the task element that is important because it is pervasive.) Some combination of these ratings can be used as the basis for selecting the task elements to be abstracted that will generalize most handily to the domain, at least as the term is used by Wheaton et al. There remains, however, the problem of putting these elements together into a task so that performance on the artificial, abstracted task will generalize to performance on the real job.

At the outset, it should be clear that this poses an empirical, not simply a rational problem. If little abstraction is involved, there may be no reasonable basis to question the relevance of the abstraction to the job as a whole; if the abstraction is severe, however, the relationship between performance on it and performance on the job itself is subject to investigation by empirical study.

Some things can be done in the test development stage, however, that will either minimize the need to do the empirical study or maximize its probability of success. These are things that maximize

the basic similarities between the real tasks and the artificial abstracted tasks.

The first report of this series, the taxonomy of testing, reported nine possible categories of task variables. These included;

1. Duration or intensity of attention
2. Hazards
3. Degree of task structure
4. Organizational involvement
5. Task complexity
6. Intrinsic feedback
7. Skill demands
8. Significance
9. Autonomy

It would appear intuitively that attempts to match the two tasks on the most salient of these kinds of variables would assure generalizability of the abstraction to the domain.

INDUSTRY-WIDE RESEARCH

There seems to be growing concern for developing research programs to validate employment tests on an industry-wide basis; the banking industry, the insurance industry, and the electric power industry have all initiated such research. In such studies, the purpose is to develop tests that will generalize across various organizations which, independently, make up the industry. In part, this

discussion belongs under the next major heading, the generalizability of relationships, since the major target is to develop criterion-related validity statements that will generalize from one company to another. However, before that portion of the objective can be reached, it is first necessary to define the job domains to be tested. Different jobs in different organizations may involve the same total collection of component tasks put together in different ways. The problem is to identify a subset of component tasks which will themselves have generality, which is to say, will generalize, across organizations.

GENERALIZABILITY OF RELATIONSHIPS

Every criterion-related hypothesis is an example of a generalization; the hypothesis, if supported, means that inferences about a criterion can be drawn (or generalized) from performance on a predictor. If the hypothesis is supported often enough, and most of the time, the validity of the hypothesis is itself generalized to situations beyond that in which it was initially supported.

THE CRITERION-RELATED HYPOTHESIS

The most commonly proposed hypothesis in personnel testing is the predictive hypothesis that performance on an employment test can be used in a functional equation to predict performance on the job. Variations on the theme include the hypothesis that a work sample can be used, perhaps in something like a discriminant function, to classify examinees correctly into master and non-master categories defined on some basis other than the test. In such examples, it is clear that the predictor, the test score variable, and the criterion, a job performance variable, are measuring quite different classes of

attributes. The fact of different variables is what sets the investigation of the criterion-related hypothesis apart from investigations of test validity which inquire into the excellence of the test as a measure of some variable.

The list of statements at the outset in this report refers to the conventional criterion-related hypothesis in statement 5; perhaps this statement describes the essential criterion-related hypothesis for personnel testing: the hypothesis that the ordering of people according to their test scores can generalize to their order according to a variable "of greater institutional concern." The essence of the criterion is that it is the variable of greater organizational interest. Whenever scores on a test are used not as descriptions of test performance but as indicators of something else that is more important, a criterion-related hypothesis is at least implied. Two kinds of implied hypotheses need to be identified, the hypothesis implied by using one variable as a presumed measure of a different variable, and the specific example of that presumption in work sample testing where an abstraction of the job is a presumed measure of performance on the real thing.

Inferences from Literal Measures. Psychological measurement typically measures one thing for the purposes of inferring something of greater interest. The number of arithmetic problems correctly solved in a specified time period is taken as a sign of an underlying construct called numerical fluency. Correct identification of where small areas on a flat surface would be found if the surface were folded into three dimensional objects is used at one level as a measure of an underlying ability called visualization and, at another, of a construct called mechanical aptitude. The example used earlier is of measuring the electrical resistance on the surface of the skin

and, as it changes, inferring changes in emotionality.

These are criterion-related hypotheses, but they are not among the examples of criterion-related validation in conventional approaches to personnel testing. Rather, the evaluation of these hypotheses proceeds along the lines known as construct validation; if a substantial body of research, consisting largely of concurrent criterion-related validity studies testing these hypotheses, supports these presumptions, then the literal measure is said to be valid for measuring (or, more properly, inferring) the more interesting measure. In a dictionary sense of the term, such an inference is a generalization from the narrowly defined literal measure to the broader meaning of the variable of greater interest. Such generalization nearly always requires some form of empirical support, although it is not always known as construct validity. In experimental psychology, particular measures are generalized because of the weight of the literature using them. If a set of operations for defining a construct becomes widely enough accepted, the issue of construct validity in its psychometric sense is not raised. Thus, if there is enough acceptance of performance on a specific set of tasks, none of which is genuinely drawn from a job, as an indication of performance on a job these tasks were designed to reflect, the construct validity of the interpretation of the performance as a general ability is not questioned.

Abstracted Work Samples. That is to say, the abstraction that occurs in the development of a work sample test may be accepted without raising questions of construct validity. The requirements are that informed people understand the nature of the abstraction and its relevance to the actual job or that a history of its use has demonstrated empirically that relevance. The latter is unlikely; unlike

experiments in basic psychology, the development of an abstracted work sample is, unless the level of abstraction is extreme, an ad hoc affair.

The nature of the abstraction and of its relevance may be self-evident to informed people if the degree of abstraction is not great. For example, it is unlikely that any informed and reasonable critic would object to a work sample test for cabinet makers that resulted in no useful product but did manage to incorporate nearly all of the most difficult kinds of joining, routing, and other skills. Real questions could be expected, however, if a test of steadiness in running a piece through the same motions used in rabbeting were used in place of cutting an actual rabbet. Where these questions arise, the developer of the test has, knowingly or not, proposed a criterion-related hypothesis in which the criterion is actual job performance. The purpose of the test is to generalize to inferences about the construct of job skill. Its construct validity may be supported in part by the logic of the content sampling leading to the abstraction, but it is better supported by empirical evidence that the generalization can legitimately be made.

VALIDITY GENERALIZATION

Assuming that the criterion-related generalization has been empirically supported, the next question that arises is whether that support holds only in specific settings. This is the question of validity generalization. Lawshe (1952) spoke of generalizable criterion-related validities as being those, like typing tests, which had found so much support that situational validity studies were not needed.

Conventional wisdom has long asserted that criterion-related validities need to be determined independently in each situation of application. The wide spread of obtained validity coefficients, such as those described by Ghiselli (1966) was taken as evidence of the need to determine for each situation the validity of the hypothesis in that situation since validities vary so widely. The present writer has noticed an embarrassing example of the hypocrisy of the demand for situational validation. In one paragraph (Guion, 1965, p. 455), it was pointed out that personnel testers have no scientific generalizations on which to depend. The next paragraph identified as the first of three "serious flaws" in selection research the tendency of personnel testers to stick with the "safe" tests -- those for which validity had been repeatedly demonstrated.

The Schmidt and Hunter Study. Schmidt and Hunter (1977) noted that many factors may be responsible for obtaining different validity coefficients in specific samples drawn from a common population. If, for example, the population consists of all people who work at a specific job tested by a specific type of test, then there is a true population validity coefficient. Obtained validity coefficients from finite (and usually rather small) samples will be distributed around that value. By making two assumptions about criterion reliability and about degree of restriction of range in these sample -- assumptions which seem not to be unreasonable -- Schmidt and Hunter determined the probable standard deviation of the sampling distribution. Their conclusion was that the variability of validity coefficients reported in various reviews can be almost entirely explained in terms of these two statistical characteristics of the samples used. When one adds further artifactual influences on obtained validity coefficients, such as differences in factor structure of individual tests within the test type or of criterion measures in individual studies, the room for

real situational variables accounting for differences in obtained validity coefficients is slim indeed.

Limits of Generalizability. The good news according to Schmidt and Hunter should not be prematurely embraced as evidence of validity generalization as the invariant rule. The logic of much of the field of industrial and organizational psychology argues against such an interpretation. Attempts to improve training will, for example, vary widely from one organization to another so that the effectiveness of job training will account for different portions of criterion variance. Organizational development programs, human relations training for supervisors, and attempts to provide or manipulate reward systems are all psychological attempts to account for substantial portions of performance variables, and the existence and success of such attempts will vary widely. Nonpsychological influences on performance should account for varying portions of the criterion variance; equipment in some organizations is the most advanced and sophisticated available; in others, it is said to be held together with chewing gum and baling wire. It is not reasonable to assume that the same "true" validity coefficient applies to the organization with the best leadership, training, and equipment as to the organization where no advances have been made in these fields.

Yet the argument of the Schmidt and Hunter analysis is persuasive; evidence can be found of the generalizability of observed relationships between predictors and the criteria they predict.

The resolution of the apparent contradiction seems obvious enough; the question is not whether relationships generalize beyond the situation in which they are first observed, but rather how far -- within what limits -- may that generalizability be assumed.

This question has not been addressed in the literature reviewed to date. Work in progress under a Fellowship at Bowling Green State University may point to one approach to it. It is currently concerned with the generalizability of various relationships at the management level. As preliminary research, three taxonomies are being developed: (a) a taxonomy of management occupations, (b) a taxonomy of management styles, and (c) a taxonomy of situations. As the work progresses, it will try out certain relationships rather well established in the research literature in each of several of the cells of the resulting three-dimensional matrix. If variability is (a) random and (b) explainable in terms of artifacts such as those described by Schmidt and Hunter, then generalizability across the domain may be assumed. If variability is (a) random and (b) greater than can be explained by statistical artifacts, then influences on validity exist which have not been included in this model. If variability is, however, substantial and fitting a pattern, then the three-dimensional model may be useful in identifying the reasonable limits within which the relationship may be said to generalize.

The taxonomic approach has serious flaws, not the least of which is the problem of arranging the cells for identifying patterns of obtained coefficients. It is, however, one approach to the search for limits of generalizability, and others need to be generated.

GENERALIZABILITY OF SCORES

The most completely described approach to generalizability is that of Cronbach, Gleser, Nanda, and Rajaratnam (1972) applying analysis of variance research designs to the study of the dependability of scores. Although the method and theory so far promulgated refer only to the generalizability of scores, they may have implications for

modifications that can help identify the limits of validity generalization as well.

Generalizability theory extends classical psychometric theory in several ways (Brennan, 1977; Cronbach et al., 1972). The most important of these is that generalizability theory recognizes distinct sources of error in any measurement, in contrast to a single, undifferentiated source of error. In classical test theory, reliability is defined as the proportion of observed score variance attributable to "true" score variance (that is, not random error variance). Generalizability theory replaces the reliability coefficient with the coefficient of generalizability, the "true" score with the more precisely delimited universe score, and random error variance in general with specific sources of error variance.

The definition of a universe depends on how the investigator plans to interpret the measure. Cronbach et al. (1972) identify three possible kinds of decisions; classification of people on the basis of scores into two or more categories, comparisons of possible courses of action for the same persons, or normative comparisons of people. The defined universe of generalization depends pretty much on the kinds of decisions. For some of these it may be the entire universe of admissible observations. For others, it may be a subset. A third possibility, that it may be something outside of that universe, has been rejected by Cronbach and his associates as either referring to validity or to an overgeneralization.

The universe score is the expected value of the observed score for a person over the universe of items. The coefficient of generalizability is the ratio of universe score variance to expected observed score variance. Clearly, there may be several coefficients of

generalizability; each of these generalizability coefficients explicates the domain to which the coefficients of generalizability are applicable.

Perhaps the major advantage of generalizability theory is that it makes the investigator ask questions which might otherwise not be considered (Cronbach, 1976). That is, the investigator must make an explicit choice of the nature of the universe to which scores are expected to generalize. The system then dictates, at least in large measure, the plan or design for data collection and analysis. The proper design depends on the nature of the universe chosen.

Replacing "true score" with "universe score" emphasizes that the investigator is making inferences from a sample of possible observations; the choice of universe emphasizes that there is more than one universe to which an investigator might wish to generalize. To be more explicit, if a person is tested on a work sample, he may himself be considered as belonging to a population of males, Californians, whites, or others; he may be tested in a population of circumstances, involving different samples of tasks, or time of day, or level of environmental hostility, or by different observers. Thus, in generalizability theory, any measurement of an attribute is considered to be a sample from some large set or universe of possible measurements, the universe of admissible observations, defined by observations in all possible combinations of circumstances.

Circumstances of the same kind are called facets. ("Facets" is chosen in preference to "factors" -- a more common term in analysis of variance -- to avoid confusion among testers with the factors of factor analysis research.) Within the universe of admissible observations, each facet consists of two or more categories or conditions.

Facets in the example above might include observers or task difficulty levels or test environments; still another might be race of the person. A decision maker may wish to generalize the results of measurement to only a limited portion of the overall universe defined by these facets; that portion is the universe of generalization. For example, a study of supervisory ratings might (conceivably) include as facets time of day of rating, nature of assignments given to the subordinate who is to be rated, race of either supervisor or subordinate or both, and height of supervisor. For particular problems, a decision made might wish to generalize only across assignments and race and ignore height or time of day.

In designing a study, all facets that might influence scores should be identified. Cronbach et al. (1972) distinguished between a generalizability study (a G study) and a decision study (a D study). The G study is designed to estimate components of variance attributable to different facets among all those that might account for substantial portions of variance; the G study therefore defines the universe of admissible observations broadly. A variety of D studies might follow from this broad definition, each of which uses whatever information is relevant to the decision maker. The distinction between the two types of studies is conventional but not very useful in this discussion of work sample testing. All work sample testing is done for the purpose of making a decision (even, if the concept of decision is broad enough, when the test is to be used as the criterion in the criterion-related validity study of some other predictor). The facets that influence work sample scores are those that indeed influence the decisions. If, for example, an abstracted work sample does not generalize from training center to field conditions, but a direct work sample does, then the decision in determining a candidate's qualifications for the field job will be based on the latter. The

discussion that follows assumes that the universe of generalizability for general evaluation of the measurement is not appreciably different from that in making decisions either about individuals or about tests, and the distinction will not be maintained.

SOURCES OF ERROR IN WORK SAMPLES

Glaser and Klaus (1962) discussed six sources of error influencing the reliability of performance measures:

1. The test environment varies; a driver's test for licensing, for example, may be given during good weather, rain, or snow.
2. System instability may cause unreliability; fluctuations in the condition of equipment (for example, wet or dry brakes) or in the behavior of other people in the system (for example, other drivers on the same freeway) influence performance.
3. The equipment may be different; in the above example, different people drive different automobiles of different ages and different systems for shifting gears.
4. The sampling of tasks may vary. The tasks in a driver's examination in a major metropolitan area are different from those in a rural county. In one state, candidates may choose between parallel parking and driving a zig-zag obstacle course.
5. Dimensions describing the complexity of the behavior required by the task may vary; they should be consistent.
6. Examinees reactions vary from time to time. Part of this is pure random variation, but part of it may be due to differences in conditions at different times, for example, when the examinee is closely observed or when the examiner is of a different race.

Since work sample tests are often scored by the judge's evaluation, a single reliability estimate is not adequate. At the very least, one needs to evaluate the reliability of the observer and the overall test reliability (Lang, 1978). A major advantage of generalizability theory

is that one can design a study to estimate the contribution to error that each of these different sources makes and, moreover, to find a generalizability coefficient which specifically accounts for whichever source of error is of particular interest. For example, one can find a coefficient of generalizability which describes whether the evaluation of performance on a work sample generalizes over different observers, over different conditions of testing, or different specific items in the work sample. In addition, a generalizability study would help an investigator decide how many judges are necessary to get dependable evaluations, or how many conditions of testing might be necessary to obtain a satisfactory generalizability coefficient.

DESIGNS

Consider a simple design. A sample of items is presented to a sample of people. In this design, using analysis of variance terminology, items and people are crossed. This is represented by the Venn diagram in Figure 1. The classical estimate of error, $\sigma^2(\delta)$, includes only the variance due to the center section of the diagram. That is, the portion of the total variances due to persons and to items interacting. Generalizability theory introduces a different error term, $\sigma^2(\Delta)$, which incorporates the additional error due to item sampling (see Cronbach et al., 1972, p. 24). Here the person is the object of the measurement and we may wish to generalize over items. Brennan (1977) presented formulas for the computations and discussed the design in greater depth. An important distinction which he made related to the appropriateness of each of these in content-referenced tests. He pointed out that, whereas $\sigma^2(\delta)$ is a measure of the variance of relative error, $\sigma^2(\Delta)$ is a measure of the variance of absolute error and is therefore much more appropriate for use in mastery testing and other forms of content-referenced measurement.

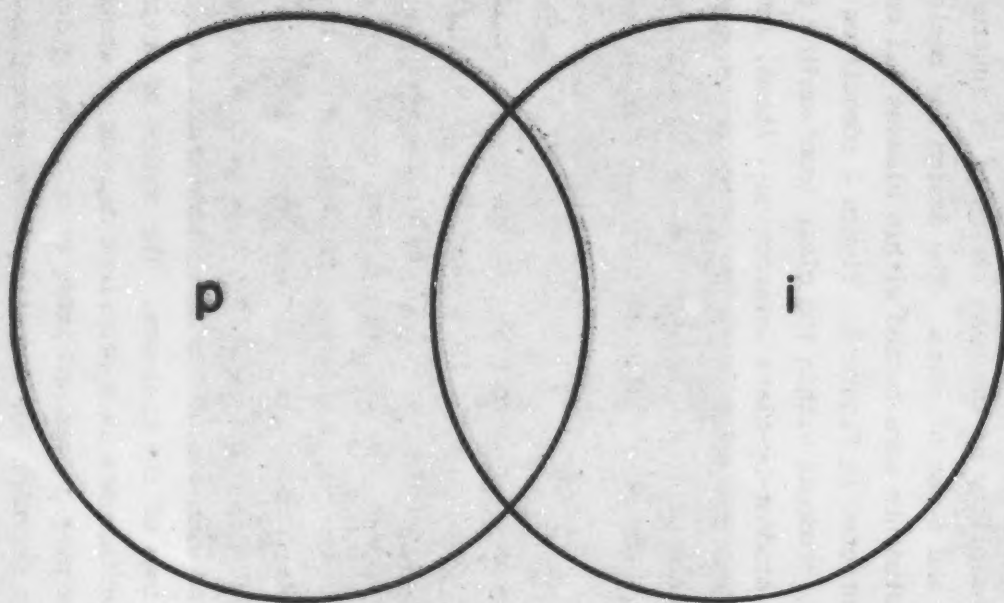


Figure 1. Venn diagram identifying variance components for estimating generalizability of scores of persons. (p = persons; i = items; area of overlap identifies p x i interaction, or error)

Dependability of Instructor Ratings. In this section, two more complicated designs which have been used in evaluating student ratings of instruction can be discussed in the context of peer ratings or observer ratings used in evaluating work samples.

The key issue in a study reported by Kane, Gillmore, and Crooks (1976) was the dependability of the mean rating of an instructor over a class of students and a set of items. The design was basically a split plot design; students were nested within classes and crossed with items, as illustrated in Figure 2. Figure 2 identifies five sources of variance: class, students within the class (confounding the student mean effect and the student-by-class interaction), items, interaction of items and class, and the residual (which includes the student-by-item interaction, the student-by-class-by-item interaction, and error). These sources of variance essentially define the nature of a generalizability study.

This may be placed in a context of peer ratings of trainee performance in a sample of work. The sample of work is done in different settings. Consider it to be a total job sample observed for a period of two days under the different conditions. The conditions might be night and day. After the two-day period, those peers who have had contact with the trainee might rate his performance on a set of rating scale items. The design for analysis is judges or peers nested within condition and crossed with items. We are interested in the dependability of peer ratings of the trainees. The choice of which of three generalizability coefficients is appropriate depends on whether one wishes to generalize over judges and items or only over judges or only over items. Explicit formulas for computing these generalizability coefficients are given by Kane et al. (1976).

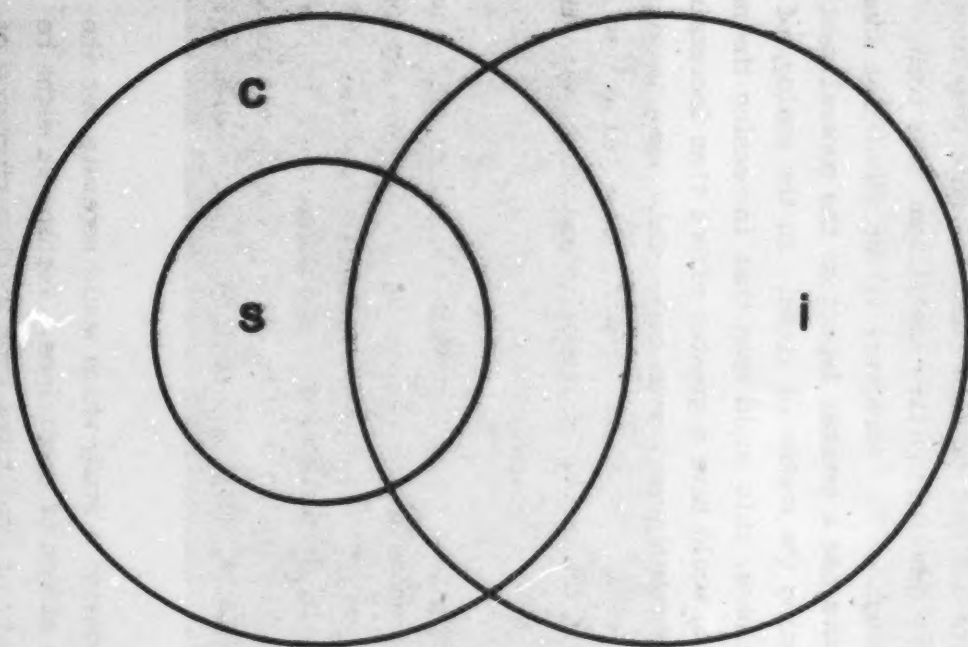


Figure 2. Venn diagram estimating variance components in student ratings of instructors. (c = class; s = students; i = items of scale; areas of overlapping circles identify various interaction terms)

A study such as this provides not only the three estimates of generalizability, but it would provide enough information so that the number of judges and the number of items could be determined for the desired size of the generalizability coefficient. In the example involving student evaluation of teaching, the authors recommended using 15 or more students and four or more items for what they considered to be a satisfactory generalizability coefficient over both students and items — roughly .70. Moreover, it was found that increasing the number of students had a greater impact on the generalizability coefficient than increasing the number of items. In the analogy of the peer ratings of trainees, this would mean that increasing the number of peers doing the rating would have a greater effect than increasing the number of items on which the rating is done. This seems intuitively logical on the basis of earlier work showing that classical reliabilities increase as a function of the number of raters in accordance with the Spearman-Brown prophecy formula.

Several questions might be asked that this generalizability design would not answer. Differences in the ratings of trainees are confounded with conditions. If one trainee is placed in a night condition and another in a day condition, differences in the ratings might be due either to differences in the trainees or to differences in the conditions. In the school example, this is a confounding of teacher and course effects.

It is possible to design a study which would separate out the trainee effect from the effect of conditions. Perhaps it might be useful to know the variety of conditions over which performance of the trainee is expected to generalize. The obvious solution is to have a number of trainees rated in each of several conditions. Such a completely crossed design may often not be feasible in practice. If

not, one might design the two studies pictured in Figure 3. In the first of these, conditions are nested within trainees, thereby providing an estimate of the variance component for the main effect due to trainees, and there is also an estimate of the variance component for the condition main effect confounded with the condition by trainee interaction. One can estimate the generalizability of the average rating of the trainee over conditions from this study. A different study, nesting trainees under condition, will give the variance component due to the main effect of conditions and an independent estimate of the variance component for the trainee main effect confounded with the trainee by condition interaction. This study will give an estimate of the generalizability or dependability of the average rating of a condition, generalizing over trainees. Computational details for this design applied in an educational setting have been reported by Gillmore, Kane, and Naccarato (1978).

For the first study in Figure 3, there are $2^3 = 8$ possible combinations of facets in the universe of admissible operations. We can generalize or not generalize over conditions, raters, and items when the object of measurement is the trainee. One can determine how the generalizability coefficients will change when the number of conditions or raters or items is changed. The logic involved in analyzing the second of these studies is similar.

Instead of conditions (night vs. day), the condition facet might have included different field conditions ranging from highly supportive, as in the training situation, to exceedingly hostile, as in realistic maneuvers or combat. A different facet might be region or part of the country or theater of operation, another one might be the nature of terrain, still another might be organizational characteristics, such as structure or "climate." In designing a generalizability study,

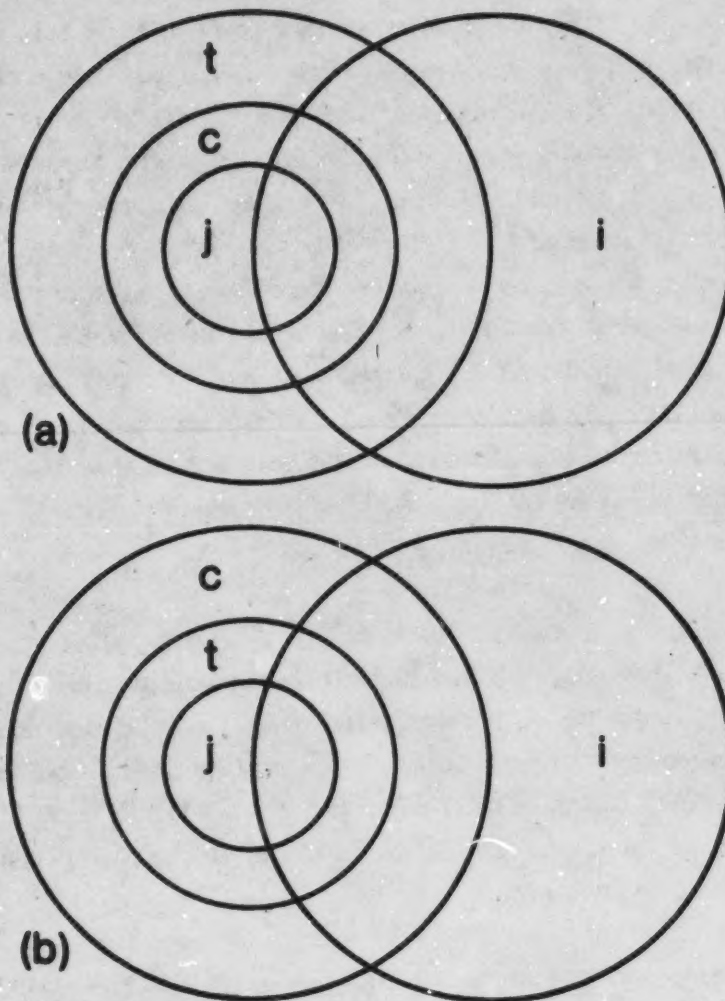


Figure 3. Two generalizability designs for estimating trainee and condition components. (t = trainees; c = conditions; j = judges or raters; i = items of rating scale)

one basically needs to identify precisely what it is that is being measured (for example, peer ratings of trainees) and the facets to which that measurement needs to be generalized. Brennan (1977) clearly covers specific details for five common designs. Cronbach et al. (1972) identifies a broader range of possible designs.

SUMMARY

The ambiguous nature of the term, generalizability, when examined leads to interesting and important research on the evaluation of work sample tests. Three meanings of the term have been identified: Generalizability of content to a broader domain, generalizability of relationships as in criterion-related validities, and generalizability of scores over conditions of measurement.

Cronbach and his associates have explicitly excluded traditional concerns for criterion-related validity from their notion of generalizability theory. A form of construct validity is demonstrated when a measure such as a work sample, which is defined primarily as a sample of job observations, is found to generalize over important varieties of conditions.

A consideration of the Schmidt and Hunter (1977) work suggests a further possibility for applying generalizability theory to the determination of the limits of generalizability in validity generalization studies. Validity generalization studies such as those conducted by Schmidt and Hunter involve the accumulation of very large distributions of validity coefficients. These validity coefficients (with the proper transformation to avoid the biased sampling distribution), can be treated as scores. The objects of measurement are not persons but studies. Facets describing the various kinds of

conditions in which validation studies are performed can be identified. An obvious example is the distinction between predictive and concurrent studies. A less obvious distinction might call for information about the subject populations and their degree of urbanization, or it might call for information about structural characteristics of the organizations. It may be that a question might arise over the generalizability of a particular validity coefficient over different industries, so that the industry of choice might be a facet. Generalizability coefficients can be computed for generalizability over various facets of interest and, by this technique, potential limits to the generalizability of a validity coefficient can be obtained. The issue of the limits of generalizability seems to be a very important one. Proper use of generalizability theory designs can lead to defined limits of the generalizability of work sample scores and of the generalizability of criterion-related validity coefficients.

REFERENCES

- Brennan, R. L. Generalizability analysis: Principles and procedures. (ATC Technical Bulletin No. 26). Iowa City: American College Testing Program, 1977.
- Cronbach, L. J. On the design of educational measures. In D. N. M. De Gruijter & L. J. Th. van der Kamp (Eds.) Advances in psychological and educational measurement. London: Wiley, 1976.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.
- Ghiselli, E. E. The validity of occupational aptitude tests. New York: Wiley, 1966.
- Gillmore, G. M., Kane, M. T., & Naccarato, R. W. The generalizability of student ratings of instruction: Estimation of the teacher and course components. Journal of Educational Measurement, 1978, 15, 1-13.
- Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.) Psychological principles in system development. New York: Holt, Rinehart, & Winston, 1962.
- Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.
- Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement, 1976, 13, 171-183.
- Lang, D. A. The work sample test: Implications for personnel research. Unpublished manuscript, 1978. (Available from Darryl Lang, Department of Psychology, Bowling Green State University, Bowling Green, Ohio, 43403.)
- Lawshe, C. H. What can industrial psychology do for small business (a symposium). 2. Employee selection. Personnel Psychology, 1952, 5, 31-34.
- Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.
- Wheaton, G. R., Fingerman, P. W., & Boycan, G. G. Development of a model tank gunnery test. (Technical Report TR-78-A24). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1978.